DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE INGENIERÍA ELECTRÓNICA, EN TELECOMUNICACIONES

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN ELECTRÓNICA Y TELECOMUNICACIONES

TEMA: DISEÑO E IMPLEMENTACIÓN DE UN SISTEMA AUTOMÁTICO PARA EL MONITORÉO DE DESASTRES NATURALES EMPLEANDO PROCESAMIENTO DE LENGUAJE NATURAL SOBRE REDES SOCIALES

AUTOR: MALDONADO ANDRADE, MIGUEL ANDRÉS

DIRECTOR: ING. ALULEMA FLORES, DARWIN

SANGOLQUÍ

2016

CERTIFICACIÓN

ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE ELECTRÓNICA Y TELECOMUNICACIONES

CERTIFICACIÓN

Certifico que el trabajo de titulación, **"DISEÑO E IMPLEMENTACIÓN DE UN SISTEMA AUTOMÁTICO PARA EL MONITOREO DE DESASTRES NATURALES EMPLEANDO PROCESAMIENTO DE LENGUAJE NATURAL SOBRE REDES SOCIALES"** realizado por el señor **MIGUEL ANDRÉS MALDONADO ANDRADE**, ha sido revisado en su totalidad y analizado por el software anti-plagio, el mismo cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, por lo tanto, me permito acreditarlo y autorizar al señor **MIGUEL ANDRÉS MALDONADO ANDRADE** para que lo sustente públicamente.

Sangolquí, 10 de Marzo del 2016

Ing. Darwin Alulema

DIRECTOR

AUTORÍA DE RESPONSABILIDAD

ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE ELECTRÓNICA Y TELECOMUNICACIONES

AUTORÍA DE RESPONSABILIDAD

Yo, **MIGUEL ANDRÉS MALDONADO ANDRADE**, con cédula de identidad Nº 1715824262 declaro que este trabajo de titulación **"DISEÑO E IMPLEMENTACIÓN DE UN SISTEMA AUTOMÁTICO PARA EL MONITOREO DE DESASTRES NATURALES EMPLEANDO PROCESAMIENTO DE LENGUAJE NATURAL SOBRE REDES SOCIALES"** ha sido desarrollado considerando los métodos de investigación existentes, así como también se ha respetado los derechos intelectuales de terceros considerándose en las citas bibliográficas.

Consecuentemente declaro que este trabajo es de mi autoría, en virtud de ello me declaro responsable del contenido, veracidad y alcance de la investigación mencionada.

Sangolquí, 10 de Marzo del 2016

MIGUEL ANDRÉS MADONADO ANDRADE

C.C.: 1715824262

AUTORIZACIÓN

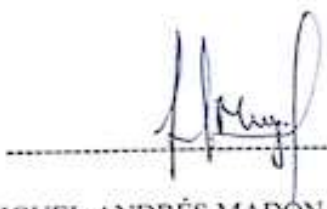
DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE ELECTRÓNICA Y TELECOMUNICACIONES

AUTORIZACIÓN

Yo, **MIGUEL ANDRÉS MALDONADO ANDRADE**, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar en la biblioteca Virtual de la institución el presente trabajo de titulación **"DISEÑO E IMPLEMENTACIÓN DE UN SISTEMA AUTOMÁTICO PARA EL MONITOREO DE DESASTRES NATURALES EMPLEANDO PROCESAMIENTO DE LENGUAJE NATURAL SOBRE REDES SOCIALES"** cuyo contenido, ideas y criterios son de mi autoría y responsabilidad.

Sangolquí, 10 de Marzo del 2016



MIGUEL ANDRÉS MADONADO ANDRADE

C.C.: 1715824262

DEDICATORIA

Principalmente a mis padres Miguel y Mery que gracias a ellos he logrado el cumplimiento de este y otros proyectos que a lo largo de mi vida me los he planteado y a mi hermano Beto por ser parte fundamental en mi vida.

Miguel Andrés Maldonado Andrade

AGRADECIMIENTO

Agradezco a mis padres por haberme enseñado los valores correctos para una persona y guiado en toda mi vida, por confiar en mí en cada uno de los pasos que he dado. A mi hermano por ser el complemento a todas las acciones que realizan mis padres.

También agradezco a mi Director de tesis, Ing. Darwin Alulema, principalmente por permitirme trabajar en uno de sus proyectos, por guiarme con sus conocimientos para la culminación de esta etapa de mi vida.

A mi familia que siempre han estado pendiente de mi a lo largo de mi vida.

A mis amigos, con los cuales compartí todos y cada uno de los momentos a lo largo de mis estudios. Gracias por su apoyo y por sus buenas intenciones.

Y a Dios por permitirme la realización de este proyecto y todas mis metas.

Miguel Andrés Maldonado Andrade

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

.

RESUMEN

Los Desastres Naturales por su naturaleza no pueden se pueden predecir con facilidad y algunos como los eventos sísmicos o volcánicos, no se ajustan a ningún modelo de predicción conocido. El Ecuador no está exento de estas situaciones y a esto se suma que se encuentra atravesado por el Cinturón de Fuego del Pacífico. Es por esto que el Gobierno Nacional ha desarrollado planes para contribuir a dar soluciones eficaces y en el menor tiempo posible, con mejoras en la infraestructura e integración de servicios. El uso de redes sociales ha contribuido para fomentar la cobertura de los hechos y brindan un seguimiento completo, ya que se encuentran globalmente difundidas por el Internet. Diariamente se generan reportes o publicaciones sobre Desastres Naturales que pueden ser analizados de distintas formas. Este trabajo presenta el diseño e implementación de un sistema automático que permite el monitoreo de la red social Twitter para filtrar el contenido en función de cuatro categorías (volcánica, telúrica, incendios y climatológica) las cuales afectan principalmente al Ecuador, y almacena todos los Tweets en una base de datos para ser analizados. El proceso de filtrado se realiza mediante el uso de la herramienta NLTK, para determinar la frecuencia de una palabra dentro de un Tweet para ser posteriormente clasificado en una de las categorías planteadas. Los resultados de cada categoría son visualizados en una página web que contiene estadísticas en tiempo real sobre la base de datos. La intención de este trabajo es facilitar el acceso a la información sobre Desastres Naturales.

PALABRAS CLAVE:

- **DESASTRE NATURAL**
- **INTELIGENCIA ARTIFICIAL**
- **PROCESAMIENTO DE LENGUAJE NATURAL**
- **TWITTER**
- **NLTK**

ABSTRACT

Natural Disasters can not be with high accuracy predicted. Some of then like seismic or volcanic events have not a specific prediction model. Ecuador is not exempt from these situations and in addition, it is part of the Pacific Ring of Fire. The Government has developed plans for providing effective assist in shortest possible time. This plans introducing improvement of infrastructure and integration services. The use of social networks has helped to coverage these events and provide full traceability because they are globally spread by the Internet. Daily reports or Natural Disasters publications can be classified for analysis. This project presents the design and implementation of an automatic system that allows the monitoring of social network Twitter to filter content based on four categories: volcanic, seismic, fire and weather. Which affect Ecuador, and stores all tweets in a database for analysis. The filtering process is performed by a tool NLTK, to determine the frequency of a word in a tweet, later to be classified in one of the categories referred. The results for each category are displayed on a web page that contains real-time statistics about the database. The intention of this work is to facilitate access to information on natural disasters, by categories depending on the kind of event.

KEYWORDS:

- **NATURAL DISASTER**
- **ARTIFICIAL INTELLIGENCE**
- **NATURAL LANGUAGE PROCESSING**
- **TWITTER**
- **NLTK**

CAPÍTULO I

1.1. Antecedentes

El Ecuador por su ubicación geográfica no está exento a catástrofes naturales como pueden ser: erupciones volcánicas, inundaciones, sequías, incendios forestales, deslizamientos de tierra, entre otros. A esto se deben sumar situaciones de riesgo que se producen en el diario vivir de la colectividad como son accidentes de tránsito, ruptura de tuberías, escape de gases nocivos para la salud, asaltos, robos, etc. Muchos de los cuales pasan desapercibidos por las instituciones de socorro, y estas se dan a conocer por personas del lugar, generando alertas por medio de las redes sociales; haciendo que los tiempos de respuesta se puedan incrementar, pudiendo llegar a hechos lamentables.

A nivel nacional la toma de decisiones sobre como actuar en el caso de una situación como las nombradas, es el Ministerio Coordinador de Seguridad (Ministerio Cordinador de Seguridad, 2014), que en función de la gravedad del hecho recurre a la Secretaría Nacional de Gestión de Riesgos y toma la acción que se debe ejecutar (Secretaría de Gestión de Riesgos , 2014). Si el hecho es local la Policía Nacional en conjunto con el Municipio son los encargados de brindar una solución.

El Gobierno Nacional en su plan de mejoras de infraestructura sobre seguridad ciudadana integró a las principales instituciones relacionadas con este tema a nivel nacional, regional y local en el sistema integrado de seguridad ECU911. Para el caso de la ciudad de Quito el sistema registró en el año 2013 los siguientes porcentajes parciales de atención: accidentes de tránsito con el 11%, requerimiento de bomberos con el 2% y manejo de gestión de riesgos con el 0,05%; separando de la estadística servicios municipales y de seguridad por parte de la policía (Ministerio Cordinador de Seguridad, 2013), esto solo para los que son registrados por las cámaras del sistema o por los agentes encargados.

Por otro lado, la cobertura de la prensa contribuye a la generación de alertas o ayudas a los ciudadanos para prevenir que sufran algún inconveniente relacionado a la situación generada. Esto ha venido de la mano del uso de las redes sociales como una rápida contribución a la cobertura por parte de las entidades de seguridad.

En el Ecuador existen proyectos para el desarrollo de sistemas de alerta temprana, como por ejemplo (Dirección Nacional de Defensa Civil, 2015), en el cual se plantean la creación de alertas ante posibles erupciones volcánicas del Tungurahua y Cotopaxi, integrando a todos los equipos de socorro que hasta esa fecha (año 2005) debían dar respuesta ante un evento como estos, pero no se han ejecutado en su totalidad, generando vulnerabilidad frente a un desastre. En el año 2010 el Gobierno Nacional puso en marcha un proyecto en la provincia de El Oro para alertar a la población fronteriza sobre las crecidas del río Zarumilla y así poder salvar vidas (Secretaría de Gestión de Riesgos, 2010).

Conjuntamente con países Sudamericanos (Colombia, Ecuador, Perú y Chile) se ha desarrollado un sistema de alerta temprano ante una amenaza de Tsunami en la región (UNESCO, 2011), difundiendo los avances constantemente en los distintos países. Actualmente en el Ecuador debido a los eventos generados por el volcán Cotopaxi, se están realizando estudios para la creación e implementación de un sistema de alerta temprana en los sectores que se verían afectados en una posible erupción.

1.2. Justificación e Importancia

La Inteligencia Artificial es considerada una ciencia relativamente nueva, esto debido a los avances de Allan Turing que la introdujo a la sociedad mediante la publicación de un artículo (Elguea, 1987). Actualmente se han centrado las investigaciones en tres postulados, el primero de estos es reconocer que el pensamiento puede ocurrir fuera del cerebro humano, es decir, en una máquina; el segundo, que el pensamiento puede ser comprendido formal y científicamente; el tercero, que la mejor forma de entender el conocimiento puede ser a través de computadoras. Lo que ha generado un sinnúmero de investigaciones, creando nuevas áreas para el estudio.

El campo de Procesamiento de Lenguaje Natural, ha generado un gran crecimiento en los últimos años (Hernández & Gómez, 2013); sus metas de estudio se han centrado en la recuperación y extracción de información para generar un conocimiento más amplio a partir de esta.

Siendo el lenguaje la forma más antigua de generar el conocimiento los humanos la han usado y perfeccionado en función de su ubicación geográfica al igual que su idiosincrasia. Sobre esto se han hecho varios estudios y clasificaciones (Hernández & Gómez, 2013). Logrando que se convierta en una área de ciencia para determinar la necesidad de la información que un usuario requiere (Ferro, 2005).

Actualmente existen lenguajes de programación, programas informáticos, técnicas estadísticas, modelos probabilísticos y algoritmos que contribuyen con procesos para la correcta extracción de datos e información para su posterior representación (Alberich, 2007).

Dando paso a modelos existentes planteados ya desde los años 2005 y 2008 (Otero & González, 2000); en los cuales se basan en la teoría de conjuntos, álgebra Booleana o el teorema de Bayes para recopilar información, se han desarrollado diccionarios capaces de facilitar la ayuda mediante herramientas de terceros, como es el caso de (Rodríguez & Carretero, 2010) un diccionario de español con cerca de 53000 términos con actualizaciones periódicas una vez comprobadas su correcto funcionamiento. Herramienta desarrollada desde el año 1994 y que su última versión estable se registra en noviembre del año 2010.

Las redes sociales hoy en día forman parte del diario vivir de las personas y su necesidad de estar conectados, siendo Facebook la que más usuarios posee, seguida de YouTube y Twitter, esta estadística viene en función del número de visitas que registran por día, al igual que usuarios registrados en las mencionadas redes sociales (Clarke & Monstesinos, 2015). Twitter en los últimos años ha evolucionado; cambiado de un servicio de microblog, y convirtiéndose en un informante de sucesos y noticias a nivel mundial, siendo la herramienta de información preferida por los usuarios, razón

por la cual esta aplicación se ha popularizado.

La expansión ha sido tan grande que las unidades de seguridad de los gobiernos, municipios, empresas públicas, etc., informan sobre desastres naturales, accidentes, realización de obras, suspensión y mejoramiento de servicios, horarios de atención, informes, precauciones que se deben tomar, etc., mediante esta red social. Pero esto no siempre es oportuno por parte de estas entidades, debido a que no se dispone de personal las veinticuatro horas del día, los siete días de la semana para que puedan estar comunicando sobre sucesos que se escapan de sus manos, y es el común de la sociedad que por este medio comunica o informa sobre desastres naturales, accidentes o problemas sociales que se presentan en el diario vivir de la sociedad ecuatoriana, creando así una comunidad que vive colaborando como informantes.

Según (Hernández & Gómez, 2013) se ha incrementado considerablemente la información digital, ya que, ésta actualmente ya no solo se genera en un computador, sino también en todos y cada uno de los dispositivos que tienen acceso a Internet pudiendo ser terminales móviles o no y con una gran tendencia a incrementar esta cantidad, más conocido como Big Data; donde se puede dividir en cinco grandes categorías estos datos: biométricos, de máquina a máquina, comercio electrónico, transacciones de datos y los generados por personas. En esta última categoría se encuentran acciones como el envío de correo electrónico, mensajes entre servicios de mensajería instantánea, postear contenidos en redes sociales, entre otros; los cuales generan una gran cantidad de datos que pueden ser analizados para brindar soluciones desaprovechando la información si no existe una herramienta que realice este trabajo.

1.3. Alcance

El proyecto comprende un diseño e implementación de un sistema automático para la realización del monitoreo de desastres naturales (erupciones volcánicas, Tsunami, movimiento de tierras) mediante el uso del Procesamiento de Lenguaje Natural empleando los Tweets generados en la red social Twitter; por lo cual, se diseña una aplicación escrita en Python que interactúa con el API de la red social, para extraer del

"Timeline" (pantalla principal de Twitter en la que se despliegan las publicaciones de las personas que se sigue) la información relacionada a desastres naturales para ser procesada con la herramienta NLTK y generar una base de datos con relación a la información obtenida.

Una vez que se extrae la información de Twitter se genera alertas para advertir sobre el tipo de emergencia que está ocurriendo, la fecha del mismo al igual que la zona geográfica o lugar donde se suscita el hecho en función de los "Trending Topics" (términos, palabras o frases más usadas) almacenados en la base de datos.

Por último se diseña una interfaz gráfica de usuario (GUI), la que permite visualizar los resultados del evento.

1.4. Objetivos

1.4.1. General

- Diseñar e implementar un sistema automático para el monitoreo de desastres naturales empleando procesamiento de lenguaje natural sobre redes sociales.

1.4.2. Específicos

- Realizar el estudio del estado del arte del procesamiento del lenguaje natural para el análisis de eventos.
- Investigar las herramientas para la implementación del procesamiento del lenguaje natural.
- Diseñar una base de datos para la recolección de Tweets para el seguimiento de eventos sociales.
- Diseñar una interfaz gráfica que permita visualizar los resultados.
- Generar reportes de los sucesos más importantes para el establecimiento de las zonas de análisis.

1.5. Estado del Arte

El Gobierno Español se ha convertido en los últimos años en un precursor de las investigaciones sobre el desarrollo de la Inteligencia Artificial y sus aplicativos para el idioma español. Ya en los años 90 las Universidades españolas enfocaban sus esfuerzos en desarrollar traductores automáticos para convertir principalmente, un texto en inglés al español, basándose en algoritmos de comparación (palabra por palabra) que poco a poco fueron evolucionando (Gelbukh, Procesamiento de Lenguaje Natural y sus Aplicaciones , 2010), partiendo de traducciones literales a unas que contengan el contexto de la frase original.

Las anteriores no son las únicas áreas donde se ha deseado desarrollar, como es el caso de los grupos Anpro21 y Brand Brian en un reporte realizado en el año 2014 (Trabazos, Suárez, Bori, & Flo, 2014) explican los avances que han realizado en base a la Inteligencia Artificial como son: un sistema capaz de analizar e interpretar imágenes, videos y voz con el fin de determinar su significado, y la temática que se está empleando es de técnicas del reconocimiento de patrones, modulación de voz y lenguaje natural; y, en el cálculo de reputación y análisis de sentimientos, donde se realiza un estudio del contexto y de la semántica sobre las noticias al igual que patrones lingüísticos sobre una persona o empresa. Pero también existen grandes contribuciones a la ciencia y específicamente en la biología como es el caso de (Glez, 2010) sobre la clasificación del ADN.

Se podría estimar que las aplicaciones que se han desarrollado en base al Procesamiento de Lenguaje Natural han crecido exponencialmente en la última década, si bien, principalmente se encuentran desarrolladas en idioma inglés, han visto la necesidad de expandirse a otros idiomas por la divulgación que el internet puede permitirlas, y de la mano de la tecnología esto a sido posible, como es el caso de Google (Google-Inc., 2015), que en el año 2000 lanzó su motor de búsqueda en más de 10 idiomas incluyendo el español, apenas tres años de su lanzamiento oficial realizado como un buscador para el inglés.

Para el Español al mismo tiempo que se realizaban los estudios sobre la IA comenzaron a desarrollarse los primeros diccionarios (Rodríguez & Carretero, 2010),

que poco a poco han ido ganando una presencia en el desarrollo de aplicaciones basadas en el Procesamiento de Lenguaje Natural en la lengua española. Comenzaron a existir trabajos que ya incluían un procesamiento más avanzado con un desarrollo e implementación de algoritmos para ampliar esta rama como es el caso de (Pérez, 1996) que gestionaba los diccionarios para poder realizar una comparación de sinónimos y antónimos de una forma gramatical más correcta y no tan robotizada a modo de una aplicación interactiva.

Otros avances relevantes han sido los estudios realizados sobre el etiquetado gramatical el momento de realizar el análisis de un texto como lo plantea (Graña, 2002), en la cual el principal objetivo se ha convertido en eliminar la ambigüedad que existe en el lenguaje. Sobre el análisis de sentimientos existen varios trabajos uno de estos es el realizado por (Dubiau, 2013) donde se plantea el uso de varios algoritmos probabilísticos para la clasificación de los sentimientos en positivos y negativos, y las diferentes herramientas que se pueden emplear; al igual que la clasificación de textos, que busca patrones para categorizar o etiquetar un documentos. Acerca de redes sociales y en especial de Twitter en Español se puede nombrar el trabajo realizado por (Alcázar, 2013) en el que realiza una clasificación de los Tweets en función de su objetividad, subjetividad, el idioma y la estructura del mismo, esto es, decir en función de la terminología de la red social y los caracteres especiales que utiliza.

En Noviembre de 2015, Google abrió la versión Open Source su motor de Inteligencia Artificial, TensorFlow, que es una librería de código abierto para computación usando flujos de datos gráficos. Su flexibilidad permite que pueda implementarse con su API en cualquier dispositivo como: servidores, computadores de escritorio o dispositivos móviles (TensorFlow, 2015). Su principal objetivo fue facilitar el aprendizaje de máquina y la búsqueda dentro de redes neuronales, pero actualmente esta librería es aplicable a cualquier área de conocimiento debido a su robustez. Escrito en Python pero con interfaces para ser usado en diferentes lenguajes.

Con esta herramienta y sus algoritmos se abre un nuevo mundo para el desarrollo de la Inteligencia Artificial en todos los campos y áreas.

CAPÍTULO 2

2.1. Introducción

Con el avance de la tecnología y la capacidad de procesamiento de las máquinas se han desarrollado diferentes herramientas que trabajan sobre el Procesamiento de Lenguaje Natural. En un principio estaban enfocadas a la creación de diccionarios sobre un idioma en específico, como por ejemplo, para: el inglés (University, 1983). En el año de 1983 ya existía un diccionario de gramática y palabras en este idioma, por otro lado, para el español existen otros diccionarios que han crecido con mucha colaboración, como por ejemplo: (Rodríguez & Carretero, 2010), que supera las once versiones y revisiones sobre sus diccionarios. Existiendo muchas limitaciones en los modelos morfológicos y sintácticos ya que para la época eran muy poco evolucionados (Román, García, & Rueda, 2012) y la computación lingüística era muy reducida. En un principio los diccionarios comparaban palabra por palabra mediante una serie de algoritmos, que clasificaban palabras en función de su entorno gramatical y así se podía llegar a determinar el idioma en el que se encontraba el texto (Alemany, 2005).

Con la evolución de las computadoras, el hombre ha querido automatizar las acciones y su diario vivir, deseando llevar a las máquinas a simular su comportamiento como son el caso de las contestadoras telefónicas; los procesadores de texto que han evolucionado hasta ser capaces de soportar funciones de dictado y realizar síntesis de texto, en función de su contenido (Gelbukh, Procesamiento de lenguaje natural, 2010); esto como evolución de la computación lingüística y el desarrollo de nuevos diccionarios como también librerías las cuales son ahora las que brindan soporte a la implementación de nuevas aplicaciones en base al procesamiento de lenguaje natural.

Actualmente, el número de librerías se ha incrementado exponencialmente, debido a la gran cantidad de lenguajes de programación que existen, siendo los más populares las herramientas desarrolladas para Python (NLTK Project , 2015). La más usada es NLTK, la cual provee para el idioma inglés un sinfín de elementos para el análisis, desde la separación de palabras dentro de una oración (tokenización) hasta un análisis sobre un tema o palabra en un texto, creación de árboles explicativos del proceso de

etiquetado, etc. Pero para el idioma español, esta herramienta tiene limitaciones, la primera es sobre caracteres especiales, tildes y letras compuestas las cuales en sus primeras versiones era imposible analizar. Actualmente con la versión 3.0, esta herramienta ya cuenta con soporte para más de diez idiomas y cambió su codificación de ASCII a UTF-8 para brindar un soporte a caracteres especiales, aumentando el número de funciones soportadas al idioma español, permitiendo ya realizar trabajos sobre esta lengua.

2.2. Inteligencia Artificial

La Inteligencia Artificial busca modelar, diseñar e implementar sistemas y/o dispositivos que posean un comportamiento que se asemege a la conducta de un ser humano (Romero, Dafonte, Gómez, & Penousal, 2007; Russell & Norvig, 2004).

A la Inteligencia Artificial se le han dado dos enfoques: uno tecnológico, y otro, científico. La parte tecnológica se enfoca en desarrollar aplicaciones y sistemas informáticos en base a programación a varios niveles para la resolución de problemas en base a algoritmos que intentan asemejarse al criterio del ser humano. Por otro lado, el enfoque científico estudia el comportamiento inteligente intentando generar teorías para comprender el comportamiento de los seres racionales e inteligentes para formular guías o esquemas para la implementación en seres artificiales (Pazos & Barreiro, 1999).

Se la puede clasificar en función del área en la que se la quiera implementar (Sotomayor, 2006) como se detalla en la tabla 1.

Tabla 1.
Clasificación de la Inteligencia Artificial.

Área	Descripción
Robótica	Máquina con capacidad de realizar procesos mecánicos repetitivos que sustituyen la presencia del hombre.
Sistemas Expertos	Estudia la simulación de los procesos intelectuales del ser humano, como el control, la interpretación de datos, diagnóstico, corrección, etc.
Aprendizaje Automático	Generación de conocimientos en base a programación computacional.
Procesamiento digital de imágenes	Estudia el reconocimiento, identificación y localización de objetos.
Almacenamiento de datos	Tratamiento inteligente de la información para la determinación de propósitos específicos.
Lenguaje Natural	Estudia la forma de interacción del humano con la máquina en la lengua propia de la persona.

En función de las líneas de investigación sobre la Inteligencia Artificial y sus sub áreas se desarrollan aplicaciones como se detalla en la tabla 2 (Pazos & Barreiro, 1999; Sotomayor, 2006; Russell & Norvig, 2004).

Tabla 2.
Sub áreas de investigación de la IA y sus principales características.

Sub área	Características
Robótica	Aprendizaje de movimientos, identificación de objetos por medio de

Continúa ⟶

	visión asistida, comunicación e interacción con humanos, etc.
Medicina	Monitorización de pacientes, interpretación de imágenes médicas al igual que la realización de diagnósticos, entre otras.
Fabricación	Diseño, implementación, monitoreo y control en el proceso de fabricación.
Sistemas y Planeación	Control de fallas, software de verificación, distribución inteligente, etiquetado, generación de información.
Militar	Planeación estratégica, codificación de mensajes, reconocimiento de objetivos.
Juegos	Interpretación de acciones, detección de jugadores y habilidades, niveles de dificultad.
Procesamiento del lenguaje	Interpretación de información, comunicación asistida, reconocimiento de lenguaje, generación de conocimiento a través del habla.

2.3. Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural (PLN) se puede considerar como una sub área de la Inteligencia Artificial. Su principal objetivo es implementar mecanismos, teorías, elementos y sistemas que permitan la interacción entre los seres humanos y las máquinas por medio de su propio lenguaje, el lenguaje natural (Cañón & Correa, 2011).

Actualmente el enfoque al cual se está canalizando toda su atención y desarrollo es el análisis de las opiniones y de los sentimientos de una persona cuando habla, escribe o se refieren a un tema o marca; esto principalmente se realiza mediante la

adquisición de datos o textos que contienen la información a analizar, bajo un criterio específico, con la definición de palabras claves que principalmente son los adjetivos calificativos, que se los clasifica como positivos o negativos, dentro del texto antes del momento de analizar, mostrando resultados en función de las estadísticas obtenidas, existen desarrollados varios proyectos el realizado en español uno de estos es JAKINBIDE (VIcomtech, 2012).

También se ha desarrollado las traducciones automáticas implementadas en aplicaciones web o para teléfonos inteligentes, el reconocimiento y clasificación de entidades mediante el uso del nombre entre otros sistemas o tecnologías (Cortez, Vega, & Pariona, 2009). Pero sin dejar de parte sus objetivos iniciales que eran facilitar la comunicación de usuarios no especializados con un computador.

2.3.1. Definición

Se puede definir al lenguaje natural como la lengua usada por los seres humanos para relacionarse, comunicarse entre sí; se incluyen los modismos y frases particulares de cada región o zona. Puede ser escrito o hablado.

2.3.2. Características, esquemas y aplicaciones

Como parte de la Inteligencia Artificial y con la evolución de la misma, comparten un objetivo en común que es la manipulación del lenguaje mediante herramientas computacionales (software). Y esto se logra mediante dos formas: epistemológico, donde se define el espacio de conceptos que el programa puede aprender; y, heurístico: se definen los algoritmos para el aprendizaje (Cañón & Correa, 2011).

El PLN analiza la estructura del lenguaje en función a cinco niveles como se detalla en la tabla (Cañón & Correa, 2011; Cortez, Vega, & Pariona, 2009) a continuación:

Tabla 3.
Descripción de los niveles de lenguaje que utilita PLN.

Nivel	Descripción
Pragmático	Contempla el significado y connotación de las oraciones en función de la situación en la que fue empleada.
Morfológico	Analiza las palabras para extraer sus raíces y otros rasgos importantes de su composición, llegando a determinar sus morfemas.
Sintáctico	Realiza el estudio de cómo las palabras pueden unir las oraciones, centrando su comprensión en la estructuración que tienen las palabras en las oraciones de una forma gramatical en función del idioma o lengua.
Fonológico	Relaciona las palabras con el sonido que representan.
Semántico	Sobre el significado de las palabras y como este da un sentido a la oración.

Las principales aplicaciones del Procesamiento de Lenguaje Natural son:

- Reconocimientos de voz.
- Resolución de problemas.
- Extracción y recuperación de información.
- Traducciones automáticas.
- Compresión del lenguaje.
- Síntesis de voz.
- Corrección de textos, entre otros.

Para analizar el lenguaje existen dos técnicas:

1. Probabilísticas: en base a una referencia, conjunto de textos de referencia (corpus), con características de tipo probabilístico que contienen las distintas fases del análisis del lenguaje.
2. Lingüística Formales: son las reglas estructurales que rigen el análisis del lenguaje.

Para la recuperación de información posee tres modelos (Rodriguez Correa & Benavides, 2007; Russell & Norvig, 2004) como se detalla en la figura a continuación:

Tabla 4.
Modelos de recuperación de la información para el PLN.

Modelo	Descripción
Booleano	Es un modelo de recuperación de información que utiliza la teoría de conjuntos y el algebra booleana, se basa principalmente en un algoritmo en criterios de decisión simples y binarios, es usado principalmente para determinar si un elemento está o no contenido en los documentos o el área de análisis Su principal objetivo es la indexación de los términos debidamente identificados. Este algoritmo tiene como entrada listas ordenadas (un par) y genera una salida una lista con la mezcla de las listas de entrada. En si ordena los documentos con los números de identificación los cuales son agrupados en función de estos.
Vectorial	Utiliza la construcción de matrices de términos y documentos, donde las columnas corresponden a los términos que se incluyen en los documentos y las filas contienen los documentos que son o fueron almacenados en una base de datos. De esta forma un documento puede representarse como un vector, donde su longitud es el número total de términos

Continúa ⟶

Probabilístico	Utiliza el cálculo de las probabilidades de que un documento tenga relevancia para la consulta. En base a un criterio preestablecido, calcula la probabilidad de que en el documento o en el documento contenga la información a recuperar, esto en función del número de términos que coinciden con el criterio de búsqueda. Cada término tiene un valor independiente, y se basa en la fórmula de la probabilidad p=n/N, donde n es el número de documentos recuperables o importantes para la extracción de información y N es el total de documentos existentes sobre los que se comparó.

Los principales inconvenientes que se presentan al trabajar con Procesamiento de Lenguaje Natural son característicos de cada idioma (Contreras, 2001) algunos de estos se detallan en la tabla a continuación:

Tabla 5.
Principales problemas que se encuentran en el PLN.

Problema	Descripción
Ambigüedad	Esto se puede suscitar debido a las múltiples interpretaciones que se puede dar sobre una palabra.
Inexactitud	Incluye los errores ortográficos, los signos de puntuación y oraciones agramaticales.
Incompletitud	Construcciones elípticas, anáforas, diminutivos, uso de sinónimos y antónimos.
Imprecisión	Uso de términos relativos sin puntos de referencia específicos y uso de términos cualitativos.

2.4. Herramientas de software para el Procesamiento de Lenguaje Natural

2.4.1. Python

Es un lenguaje de programación, considerado de alto nivel, desarrollado por Guido Van Rossum en los años noventa. En el año de 1991 se lanzó la versión 0.9.0, primera versión de este lenguaje, con licencia open source (código abierto, de software distribuido y desarrollado libremente). La versión 2.0 fue lanzada en el año 2000. Actualmente para su desarrollo existe el colectivo Python Software Foundation, el cual es dirigido por Van Rossum (Roldán, 2015). En la figura a continuación se visualiza el logotipo actual de Python (Python Software Foundation, 2015).

Figura 1. Logotipo actual de Python.
Fuente: Python Software Foundation.

Los principales objetivos para el desarrollo de Python fue que sea un lenguaje de alto nivel fácil de aprender y de lectura.

Se han distribuido tres versiones de Python, actualmente se encuentran vigentes las versiones 2.7.9 y la versión 3.5, existe una manera de identificar las versiones de Python, esta es mediante tres números separados por puntos X1.X2.X3 que indican a detalle la versión en la que se encuentra el lenguaje, donde X1 indica el lanzamiento de las grandes versiones (versiones: 1, 2, 3); X2 indica la introducción de grandes novedades en las grandes versiones; y, X3 indica las correcciones de errores y fallos de seguridad, se puede considerar como lanzamientos menores. Como se puede

observar en la figura a continuación en la que se detallan las versiones de Python (Marco, 2015).

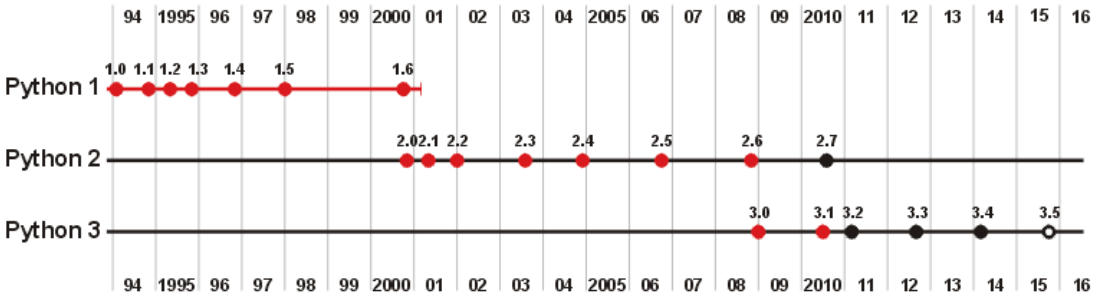


Figura 2. Versiones de Python a lo largo de su historia.
Fuente: Bartolomé Marco.

En el Anexo 1 se detalla la instalación de Python sobre los entornos Unix y Windows.

2.4.1.1. *Estructura de programación*

En comparación de otros lenguajes de programación Python no necesita una secuencia de escritura o el inicio con una estructura cerrada, pero si bien no existe una norma que implique el orden en la programación de Python, se recomienda que se lleve un esquema referenciado en comentarios que expliquen el uso de cada clase, función o variables, esto para contribuir a una facilidad en la lectura de los programas, como también en la organización del script para la reutilización de código y corrección de errores. También se debe mantener la indentación dentro de sus funciones, ya que es la única que determina el cierre de bloques (clases, funciones, estructuras de control) (Soriano, 2009). Por estas razones y por una lógica en el script, se debe llevar el siguiente esquema:

1. **Línea de iniciación:** usado solo en entornos Unix, permite la ejecución del script con su nombre solamente.
2. **Documentación:** breve descripción del script, se informa su utilidad y el objetivo.
3. **Importación de módulos**: se importan los módulos necesarios para la ejecución del programa.

4. **Declaración de variables:** se declaran las variables globales.
5. **Declaración de clases**.
6. **Declaración de funciones**.
7. **Cuerpo del programa (main):** Lugar donde todo el código se ejecuta.

Soporta varios tipos de datos como: booleano, entero (int), complejo, flotante, cadena de caracteres (string), listas, tuplas y diccionarios (hash), donde existen funciones generales que se aplican a la mayoría de tipos de datos, como también existen funciones dedicadas para el manejo específico.

2.4.2. NLTK

Natural Language Toolkit (NLTK) es una librería *open source* para Python que permite trabajar con datos de lenguaje de humano, diseñada por el curso de lingüística computacional de la Universidad de Pensilvania. Con un objetivo pedagógico, soportando: asignaciones, con soporte para edición de contenidos, modificación de métodos con el fin de poder contribuir al crecimiento de la herramienta; demostraciones, posee gráficas y enseñanza paso a paso que contribuyen al aprendizaje; y, proyectos, brinda la flexibilidad para que se puedan desarrollar proyectos a partir de cero (Bird & Loper, 2007). Está diseñado no solo para la búsqueda en el texto ni extracción de información sino también para dar un tratamiento al texto que se desea analizar, convirtiéndose en una herramienta poli funcional para el Procesamiento de Lenguaje Natural. Además de ser multiplataforma, posee una documentación extensa sobre todas las sub herramientas que posee para el procesamiento.

Dentro de la herramienta existe una serie de términos y definiciones que se deben tomar en cuenta para un correcto uso, la mayoría de estos se detalla en la figura, a continuación (NLTK Project , 2015):

Tabla 6.
Terminología usada por la herramienta NLTK.

Término	Significado
Token	Representa una palabra, es la unidad más pequeña y simple en el procesamiento.
Sentencia	Secuencia de Tokens ordenadas.
Tokenización	Proceso en el cual a una sentencia se la separa en Tokens que la componen, también verifica el contenido incluyendo los espacios y separaciones.
Corpus	Cuerpo del mensaje que se encuentra en una serie de sentencias.
POS Part-of-speech	Clasificación de cada Token dentro de una sentencia en función de su categoría gramatical.
Árbol	Forma de ordenar cada uno de los Tokens dentro del corpus de un texto.
Etiquetado POS	Proceso en el que se identifica y clasifica a cada uno de los Tokens dentro de las sentencias del texto con su correspondiente POS.
Parsers	Proceso que se encarga de formar los árboles de Tokens, para realizar esta acción se basa en el etiquetado POS.
Morfología	Proceso que extrae de los Tokens los morfemas y las raíces en función de la gramática empleada y lenguaje.

Para su funcionamiento necesita la instalación de paquetes adicionales o complementos que contienen los corpus, reglas gramaticales y otras ayudas complementarias como funciones y métodos para trabajos en texto. Los primeros pasos para su uso: son la comprobación de la instalación correcta y esto se realiza ingresando mediante Python a su propio interprete de comandos y se procede a la importación de

la librería NLTK, *import nltk*, de no existir error se puede proceder a ejecutar la función que permite la descarga de los complementos, *nltk.download()* resultado que se muestra en la figura 3. Actualmente cuenta con doce módulos que se categorizan en función del área de investigación o trabajo.

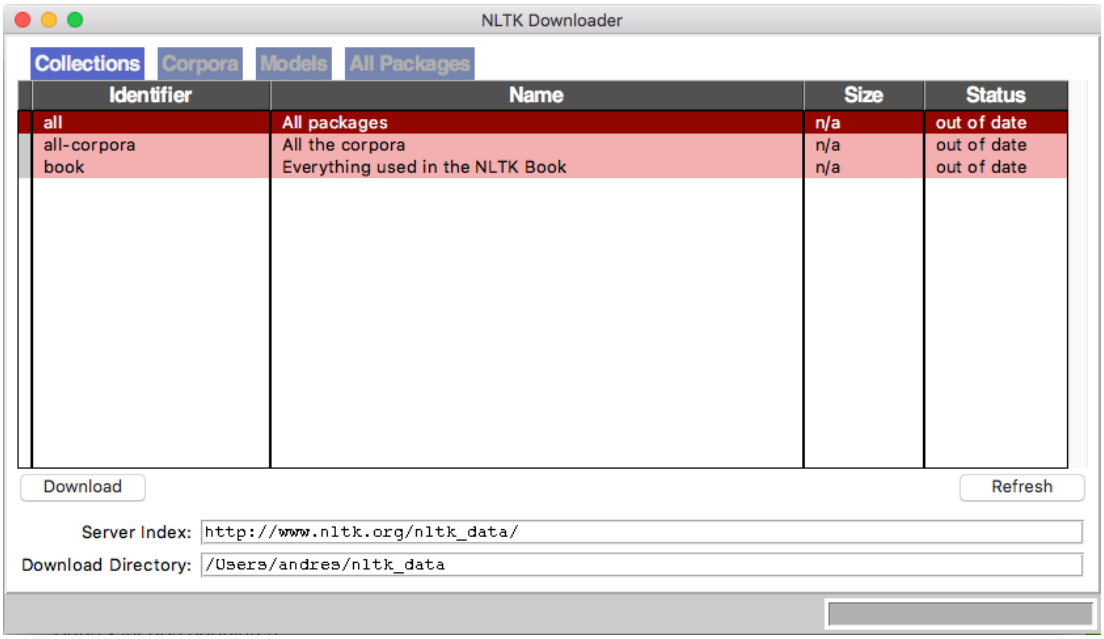


Figura 3. GUI de NLTK book collections de paquetes complementarios.

Los corpus se manejan sobre varias estructuras y estas se realizan en función del género, la fuente, el autor, el idioma en el que se encuentra el texto entre otras. Las estructuras se clasifican para NLTK en cuatro categorías que son: el aislamiento o aislar, categorización, superposición y temporal.

Después de leer o acceder a un texto, se le puede asignar a una variable para facilitar su procesamiento, la cual de preferencia trabajar con el tipo de dato string. La primera forma de trabajo que se puede realizar es la Tokenización del texto, NLTK trabaja principalmente con la codificación “ASCII” para caracteres en español se puede emplear la codificación “UTF-8”; en la que se debe decodificar el texto a procesar y codificar el texto a presentar o imprimir por terminal evitando así los problemas con caracteres especiales, luego se puede normalizar los tokens en base algún criterio necesario o requerido para agregarlos a un diccionario. Una vez

cumplido el proceso de Tokenización se puede etiquetar los tokens con su respectiva categoría gramatical (NLTK Project , 2015).

2.4.3. Twitter

Nace el 14 de marzo de 2006, en las manos de Jack Dorsey, en un período corto de programación; la idea original de los fundadores de Twitter fue desarrollar una empresa de postcast, la cual fue un fracaso ya que nunca llegó a tener un lanzamiento oficial. Se concibió como un proyecto de investigación, inspirada en la aplicación de intercambio de fotos Flickr (sitio web en el cual se permite el almacenamiento, búsqueda, venta y compartición de fotografías y videos a través de internet, de propiedad de Yahoo Inc.) y por eso su primer nombre Twttr. En el año 2009 fue introducido completamente al español (Jorge, 2011) y actualmente el diseño de su logotipo se muestra a continuación (Twitter Inc, 2015).

Figura 4. Logotipo de Twitter.
Fuente: Twitter Inc.

La idea principal de esta nueva aplicación fue desde sus inicios la misma que se ha mantenido hasta ahora, esto es, diseñar una red social que permita comunicar a sus usuarios con un límite de 140 caracteres de simple acceso y de fácil utilidad, como un sitio de microblogging que permita una interacción rápida de los usuarios con sus pensamientos, ideas, su diario vivir. Si bien no se ha dado un uso específico para la aplicación pero se ha ido diversificando, actualmente el principal uso de esta herramienta ha sido el informativo, en el que periódicos, reporteros y el común de las personas informan sucesos, hechos o eventos que pasan a su alrededor. Otro de los usos, ha sido el de promoción de productos, mediante campañas y el uso de hashtags para crear tendencias sobre lo que se quiere promocionar, fuerte herramienta para las empresas que no desean gastar grandes cantidades de dinero en campañas televisivas (Domizi, Roma, Fiadesio, Lantari, & Montesano, 2013).

Dentro de la red social existe una terminología la cual se detalla en la tabla a continuación (Martínez, 2012; Mollet, Moran, & Dunleavy, 2011):

Tabla 7.
Terminología usada en Twitter.

Término	Definición
@nombre_de_usuario	El nombre de usuario escogido es precedido por el signo @ como identificación en la red social.
Tweet	Publicaciones realizadas en la red social, posee una limitación de 140 caracteres.
Timeline	Historial de las publicaciones propias y de amigos.
Follower	Seguidor de las publicaciones de otro usuario.
Follow	Seguir las publicaciones que un usuario realiza.
@Reply	Son Tweets en el cual se menciona a uno o varios usuarios, de una forma pública.
RT o Retweet	Compartir el Tweet de otro usuario.
DM o DirectMessage	Tweet o mensaje directo enviado a un usuario de forma privada, no se visualiza en el Timeline.
# Hashtag	Se la utiliza para etiquetar una palabra o frase, con la cual se pueden clasificar los Tweets.
Favorites	Tweet marcado como favorito, o del cual se indica agrado.
Trending Topic	Palabras o frases más utilizadas, comúnmente se conoce como “lo que se está hablando en ese momento”.

2.4.3.1. *Api de Twitter- Tweepy*

Twitter ofrece un entorno para desarrolladores de fácil acceso y uso, para integración con blogs, aplicaciones móviles y otras webs. Esto lo realiza de dos maneras: la primera es una consola de desarrollo y programación dentro de la misma página web de Twitter, en la que se pueden realizar ciertas mediciones de Tweets, que implican el tráfico, manejo de trending topics y creación de alertas; la segunda opción son una serie de librerías implementadas por diferentes desarrolladores para varios lenguajes de programación como por ejemplo: Java, C++, .net, JavaScript, PHP, Ruby, Python, entre muchos otros (Twitter Inc, 2015).

Para Python la lista de librerías también es grande, pero Tweepy por su sencillez, su gran información con respecto a la documentación y de no requerir un alto manejo de Python hacen que sea ideal para la realización de este proyecto, actualmente se encuentra en la versión 3.5 que es compatible con cualquiera de las versiones disponibles y que poseen soporte en este lenguaje (2.6, 2.7, 3.3-5). La instalación de esta librería se detalla en el Anexo 3.

Divide su estructura en dos partes: el Api que es una clase que permite el acceso a todos los modelos en Twitter, donde se pasan parámetros para obtener una respuesta, el equivalente a funciones; los modelos son aquellos que se invocan en los métodos Api y estos contienen la información solicitada por parte de Twitter (Roesslein, 2009).

Los métodos pueden clasificarse en varias clases, las que contienen funciones específicas sobre el requerimiento sobre el que se desea aplicar, al igual que parámetros específicos que se utilizan dentro de sus funciones, como se detalla en la tabla a continuación:

Tabla 8.
Métodos existentes en la librería Tweepy para Twitter.

Tipo de Método	Descripción
Timeline	Posee funciones que trabajan sobre la manipulación del Time line como la extracción de Tweets, visualización de re-Tweets, entre otros, en base a parámetros como la identificación de un usuario, el número

Continúa ⟶

	de páginas de recolección o el número de Tweets que se desean obtener.
Status	Trabaja directamente con el status: obtención, actualización, Re Tweet del desarrollador y de usuarios en base a la identificación del usuario, al igual que la eliminación de Tweets.
User	Contiene funciones que trabajan con los datos del usuario, información de contacto, seguidores y búsqueda de usuarios.
Direct Message	Manejo y envío de mensajes directos entre usuarios, recolección y eliminación de los mismos.
Friendship	Con estas funciones se puede realizar un *follow* o *unfollow* dentro de los datos del usuario, al igual que visualizar los seguidores y a quien se sigue en la red social.
Account	Principalmente con estas funciones se puede dar una administración reducida de la información de contacto.
Favorites	Administración de los Tweets marcados como favoritos o indicados agrado.
Block	Contiene las funciones requeridas para bloquear a un usuario o su contenido.
Spam Reporting	Contiene una función que permite reportar a un usuario como *spammer*.
Saved Searches	Administración de contenido de usuario como son la autorización de credenciales, eliminación y modificación de las mismas.
Help	Provee consultas sobre un Tweet específico.
List	Administración para la creación de nuevas autentificaciones de usuario.
Trends	Entregan información sobre la ubicación, los temas que son trendding topic en función de una localización específica o el lugar donde se realizó la consulta.
Geo	Referencias geográficas sobre el origen de los Tweets.

Para el funcionamiento y como parte del uso de las herramientas de desarrollo de Twitter, Tweepy requiere de una autorización, que en Python son cuatro variables y estas son: *consumer_key*, *consumer_secret*, *access_token* y *access_token_secret*. Mediante el uso de una función que valida el contenido de las variables procede a acceso de los datos que están disponibles para los desarrolladores.

Posee algunas limitaciones, una de estas es en función del número de consultas a Twitter generando una ventana y si esto se supera existe un bloqueo de sesenta minutos, luego de este tiempo regresa a la ventana original. Otra limitación es el acceso al historial del Time line que puede hacerse en función de las páginas o Tweets, y este se limita a dos y doscientos respectivamente con los parámetros ingresados para el tipo de captura, todo esto en función de la versión web y su contenido.

2.4.4. Base de Datos

Puede definirse como una colección de archivos o datos que se relacionan y permiten manejar información mediante una serie de programas (Silberschatz, Korth, & Sudarshan, 2002). Donde cada archivo o dato es visto como una colección de registros y estos a su vez como una colección de campos. Los campos poseen información sobre un atributo que corresponde a una entidad real. Los archivos pueden ser vistos como tablas.

Para la administración se requiere de un sistema, estos tienen como objetivo principal la organización y la manipulación de los volúmenes de datos que se encuentran en la base de datos.

Existen varios tipos de datos que se pueden almacenar en una base de datos, los alfanuméricos y numéricos, son los principales pero también existen otros como: los lógicos, fecha, memo y general, entre otros.

Para el diseño, implementación y manejo de base de datos existen tres enfoques (Cruz, 2008), los cuales se detallan en la tabla a continuación:

Tabla 9.
Descripción de los enfoques de diseño de una base de datos.

Enfoque	Descripción
De redes	Toma la información como si fuesen conjuntos: propietarios y miembros. Se

	reduce la redundancia y no existe pérdida de información.
Jerárquico	Visualiza a la base de datos como una relación padre-hijo, genera excesiva redundancia de datos.
Relacional	Trata a los datos como un conjunto de tablas, en donde cada renglón es un registro y las columnas son los campos donde se describen estos registros

2.4.4.1. *MySQLdb*

Es una librería desarrollada para comunicar la interface de las bases de datos y servidores MySQL como un API para Python. Fue desarrollado en un principio para el lenguaje C y debido a su alta popularidad de desarrollo para varios lenguajes de programación entre ellos Python.

Contiene funciones que permiten el manejo y administración de base de datos desde un script de Python. Se realiza en principio la comunicación con la base de datos la cual parte de cuatro parámetros que son: el lugar o dirección del servidor, nombre de usuario, la base de acceso, y la base de datos en la que se va a trabajar. La dirección del servidor viene dada por defecto y es “localhost” o 127.0.0.0.1.

Para realizar una consulta se debe crear una variable que contenga la ubicación de un cursor para determinar la posición en la base de datos, al tener diferentes sintaxis Python y MySQL se debe pasarla la sentencia que contiene la acción a realizar en formato string a MySQL para que en este sea comprendido y pueda devolver el resultado de la consulta correctamente.

2.5. Herramientas complementarias para el desarrollo del proyecto

2.5.1. HTML

Sus siglas en inglés significan: *"HyperText Markup Language"* y se originó en los años 80, por Tim Berners-Lee que propuso un sistema nuevo de hipertexto para compartir los documentos. El primer documento formal bajo esta nueva característica se publicó en 1991. Pero es en 1993 que se estandarizó por parte del IEFT, organismo regulador del Internet (Martinez, 2008).

Se la puede considerar como una herramienta que sirve o se utiliza para la creación de páginas web, ya que al no necesitar ser compilado puede ejecutarse desde cualquier navegador. Uno de sus principales objetivos como lenguaje es ser reconocido mundialmente y que permita la publicación de información de una forma global. La estructura de su lenguaje se basa en etiquetas y sus respectivos atributos a manera de instrucciones para su posterior ejecución.

Las etiquetas interactúan con el equipo para que este comprenda lo que el desarrollador desea mostrar, son invisibles para el usuario final. En estas se indican la naturaleza del texto que encierran, principalmente título, imágenes, párrafos, formularios, etc. Existen dos tipos de etiquetas y estas se detallan en la tabla a continuación (Menéndez, 2014):

Tabla 10.
Características de los tipos de etiquetas en HTML.

Tipo de Etiqueta	Características
De dos en dos	Se indica su inicio, su contenido en texto y su delimitación final
Huérfanas	Sirven principalmente para la inserción de elementos en lugares específicos, no se necesita poner delimitaciones

Los atributos son las opciones que se puede aplicar a una etiqueta, son un complemento que proporcionan información adicional. Se los coloca después del nombre de la etiqueta.

2.5.2. PHP

Es un lenguaje de programación interpretado con una sintaxis similar a Java o C++, pero puede ser usado para realizar cualquier desarrollo de aplicación, pero principalmente su uso es para la generación dinámica de páginas web. Su código se incrusta en el lenguaje HTML siendo servidor web que logra ejecutarlo. Fue creado en 1995 por Rasmus Lerdorf como scripts de Perl para la realización de trabajos online con comunicaciones a bases de datos para desarrollar sitios web personales que incluían formularios y consultas (The-PHP-Group, 2015).

Sus principales características son (Palomo & Montero, 2010; Vázquez, 2008):

- Soporte para la programación orientada a objetos.
- Soporte para conexión de base de datos.
- Soporte de codificación de varios idiomas.
- Alta velocidad de ejecución ya que no utiliza muchos recursos del computador.
- Estabilidad, al poseer un sistema propio de administración.
- Varios niveles de seguridad.
- Multiplataforma.
- Extensa documentación.

CAPÍTULO 3

3.1. Diseño de la aplicación

El proyecto recolecta los Tweets del time line y los almacena en una base de datos bajo un criterio de almacenamiento y filtraje para luego ser procesados y analizados con una herramienta de Procesamiento de Lenguaje Natural. Posteriormente estos resultados son expuestos en una aplicación web que permita la interacción con el usuario, esquema que se muestra en la figura a continuación.

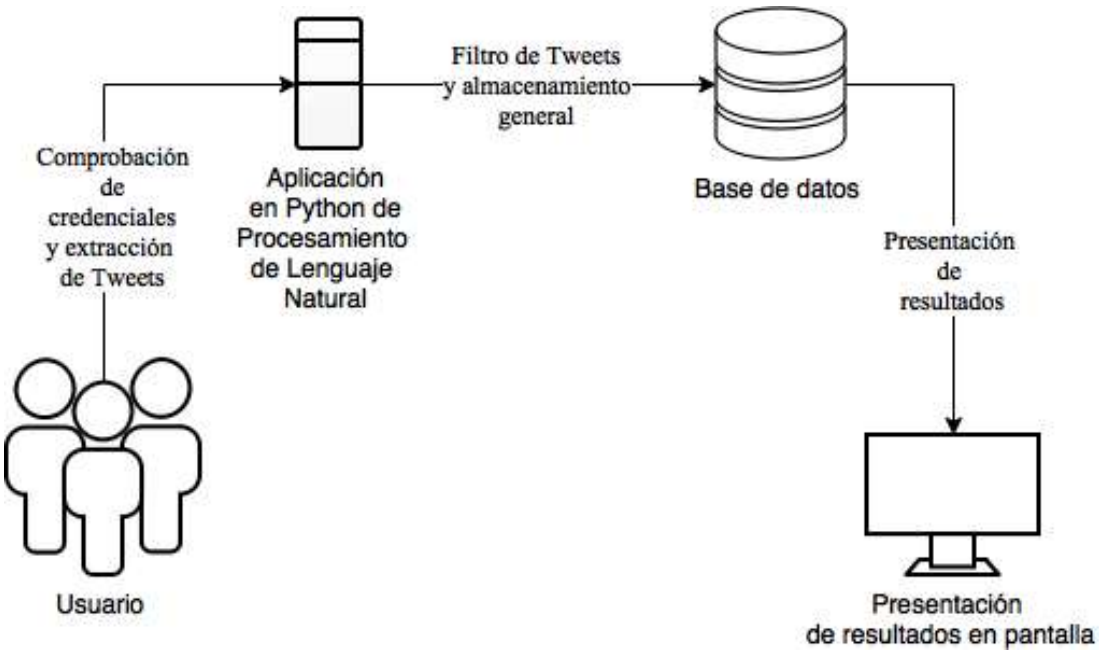


Figura 5. Esquema general del proyecto.

Previo a este capítulo se analizó las herramientas/tecnologías más convenientes para la realización del proyecto y en la tabla a continuación se resume los componentes de software que se utiliza.

Tabla 11.
Información sobre el software utilizado en el proyecto.

Componente	Fabricante / Autor	Versión
Python	Python Software Fundation	2.7.10
NLTK	NLTK Project	3.1
Tweepy	Roesslein	3.5.0
MySQL	Oracle Corporation	5.7.10
MySQLdb	Andy Dustman	1.2.5
HTML	Word Wide Web Consortium	4-01 – 5
PHP	PHP Group	5.6
XAMPP	Apache	5.6.15-1

3.1.1. Requerimientos del Sistema

3.1.1.1. *Twitter- Extracción de Tweets*

Twitter posee una limitación de 140 caracteres en sus publicaciones, aquí se incluyen URL's de direcciones Web o fotografías, también contiene información de la persona que generó el Tweet así como la fecha y la hora en la que se lo realizó. Adicionalmente en nuevas versiones de su API se incluyen métodos para obtener la ubicación en el sistema de coordenadas de latitud y longitud para la creación de un diccionario que relacione estos datos con un punto en concreto. La herramienta Tweepy permite realizar completamente estas acciones con leves limitaciones.

Dentro de la estructura de Twitter cada parámetro tiene un formato, los Tweets o el texto dentro de estos tienen un tipo de variable o formato "Unicode", la fecha y hora pertenecen al tipo de dato respectivamente "fecha y hora", el nombre de usuario es del tipo "Unicode".

El proceso de extracción de Tweets se detalla en la figura 6, donde la primera acción que se realiza es el ingreso de las credenciales (consumer_key, consumer_secret, access_token, access_token_secret) para ser comprobadas mediante

una función de autorización, si estas son correctas los métodos descargaran los Tweets para ser almacenados en una base de datos; si la verificación de credenciales es incorrecta se debe repetir el reingreso de las mismas.

Figura 6. Proceso de extracción de Tweets.

3.1.1.2. *Aplicación Python*

La aplicación en Python está dividida en un script principal y varias librerías que complementan o contribuyen a la realización de la misma. En el main o script principal se realiza el proceso de extracción de Tweets al igual que su procesamiento, es decir,

aquí se importa la librería de Tweepy y se realiza de autorización y verificación de credenciales.

Antes de almacenar los Tweets en la base de datos esos pasan por una función creada llamada Supresión, esta convierte el texto en tipo “string”, para un mejor manejo, luego realiza comparaciones sucesivas y reemplazo de los caracteres especiales (tildes, la letra “ñ” y signos de exclamación y admiración que no están disponibles en el idioma inglés), como se detalla en la tabla 12, que afectan directamente al texto en la base de datos o en el momento de realizar la ejecución de la herramienta NLTK, todo esto para estandarizar una sola forma de lenguaje y de un mismo tipo de dato.

Para finalizar elimina el URL de cada Tweet con la función “partition” se vuelve a convertir este texto en “string” y se selecciona de este nuevo vector la posición “0” que es la que contiene el texto trabajado, algoritmo que se explica en la figura 7, para luego ser asignado a una variable.

Tabla 12.
Reemplazo de caracteres en la función suprimir.

Carácter	Reemplazo	Herramienta Afectada
á	a	NLTK
Á	A	NLTK
é	e	NLTK
É	E	NLTK
í	i	NLTK
Í	I	NLTK
ó	o	NLTK
Ó	O	NLTK
ú	u	NLTK
Ú	U	NLTK
ñ	nio	NLTK
Ñ	NIO	NLTK
‘	Espacio	MySQL
(	Espacio	NLTK

Continúa ⟶

)	Espacio	NLTK
¿	?	NLTK
¡	!	NLTK

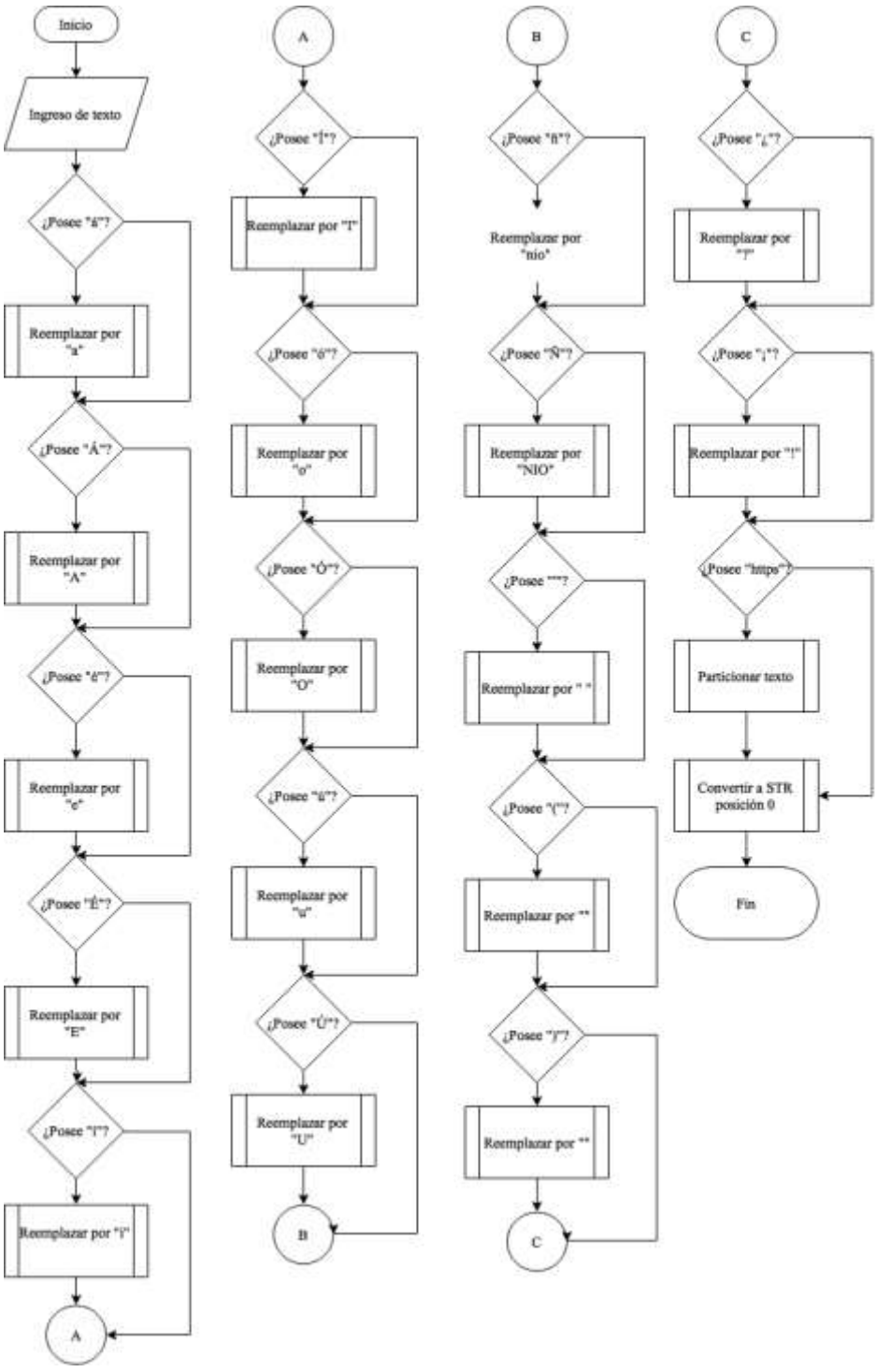


Figura 7. Algoritmo empleado para la supresión de caracteres especiales.

Para la presentación en consola se da un formato a los datos descargados de Twitter, es decir, una viñeta para indicar el inicio de cada Tweet, luego se da la fecha y la hora en la que se lo realizó seguido por el autor para continuar con el Tweet propiamente dicho, tal como se describe en el esquema de la figura a continuación. Bajo este formato se guardan completamente los Tweets con los campos mencionados en una tabla principal de la base de datos.

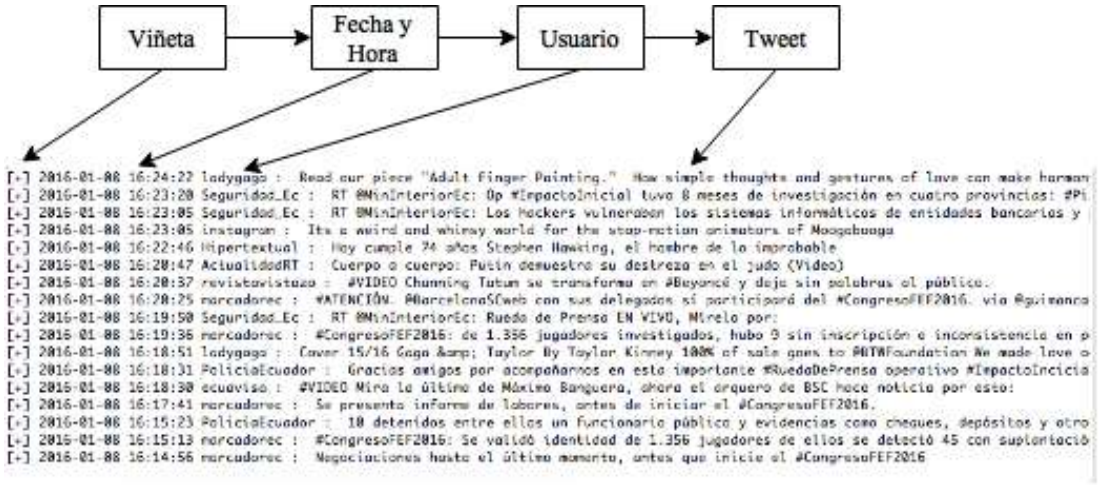


Figura 8. Esquema de presentación en consola la extracción de Tweets.

3.1.1.3. *Categorías de clasificación para el filtrado*

Para cumplir con los requerimientos de la aplicación se deben clasificar los Tweets en función del tipo de desastre natural al que corresponde, es por esto que se dividió en cuatro categorías a los desastres naturales más comunes que ocurren en el Ecuador, tal como se detalla en la tabla a continuación, y de esta forma contribuir al proceso de filtrado.

Tabla 13.
Palabras claves dentro del proceso de filtrado por categorías.

Categoría	Palabras Claves
Volcánicos	• Erupción. • Lahar. • Azufre. • Volcán. • Ceniza.
Telúricos	• Sismo. • Terremoto. • Movimiento de tierras. • Temblor. • Epicentro. • Falla geológica. • Deslave.
Incendios	• Fuego. • Conato. • Incendio. • Forestal. • Bomberos. • Incendiario.
Climatológicos	• Lluvia. • Inundación. • Tsunami. • Maremoto. • Marejada. • Aguaje. • Fenómeno del niño.

3.1.1.4. *NLTK*

Para esta sección y parte del trabajo, se centró en la realización del proceso de filtrado y el uso de la herramienta NLTK basándose principalmente en los modelos de recuperación de información: Probabilístico y Booleano, previamente ya descritos.

Este proceso se realiza en función de cada categoría definida anteriormente con la creación de un conjunto de palabras, tomando referencia el modelo Booleano ya que cada categoría contiene varias palabras relativas a su tema formando así grupos o conjuntos debidamente identificados, los cuales serán comparados.

A cada Tweet descargado del Time line se realiza el proceso de Tokenización, con la función word_tokenize(), parte de la herramienta NLTK, que divide el string obtenido después del proceso de supresión en un vector, creando una sentencia de Tokens, por cada Tweet extraído o descargado, los cuales son almacenados en una variable auxiliar para las próximas operaciones.

Seguido de esto, se crea un objeto que contendrá la información relativa a la frecuencia de una palabra dentro de la sentencia, con la función FreqDist() obtenida con la importación de la librería NLTK al script, todo esto se realiza como parte del proceso de filtrado.

A continuación se analiza cada sentencia con la función freq(), siento estas N o el total de la información a examinar, calculando la probabilidad de que en cada Tweet exista una o varias de las palabras definidas previamente en el conjunto creado para cada categoría (n) para la clasificación de los desastres naturales, si la probabilidad es mayor a cero (0) es decir existe la palabra dentro de la sentencia, se reasigna el contenido y se almacena en la tabla de la base de datos a la que corresponde en función de sus parámetros y coincidencias obtenidas en el proceso al igual que se mantiene la sentencia en la tabla principal; para los caso donde no existe una probabilidad mayor a cero (0) el Tweet se lo almacena solamente en la tabla principal. Este proceso completo se detalla en la figura 9.

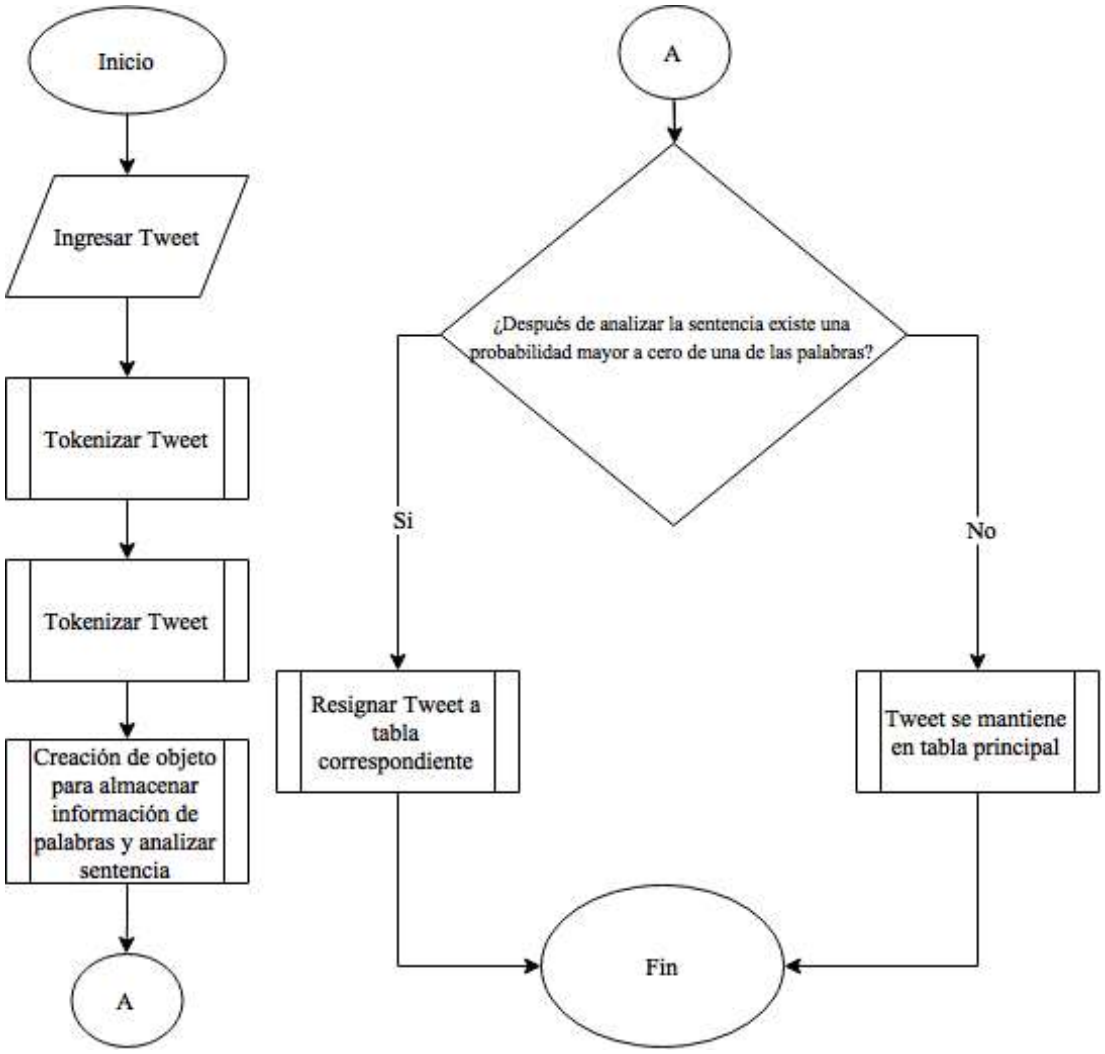


Figura 9. Algoritmo para el filtrado de Tweets.

3.1.1.5. *Base de datos*

En la base de datos se procura almacenar toda la información sin discriminar el tipo que esta pueda presentar o generar para el proyecto, por lo cual se ha dividido en varias tablas como se muestra en la tabla a continuación:

Tabla 14.
Principales características de las tablas dentro de la Base de datos.

Tabla	Descripción
Principal (Twitter2)	En esta tabla se recolectan todos los Tweets obtenidos del Time line indistintamente.
Volcán	Para esta tabla se recolectan todos los Tweets referentes a eventos volcánicos al igual que la información referente a los principales volcanes del país.
Terremoto	Esta tabla recolecta los Tweets sobre movimientos telúricos, sismos, terremotos, entre otros.
Fuego	Donde se recolectan los Tweets que hacen referencia a incendios y todo lo que con estos conlleva.
Clima	Para esta tabla se recolectan los Tweets referentes a eventos climatológicos tales como: inundaciones, lluvias, fenómeno del niño, tsunamis, aguajes, entre otros.

Cada tabla se maneja bajo un mismo esquema el cual se muestra en las figuras 24, 25, 26, 27 y 28 para cada una de las ya mencionadas.

Figura 10. Descripción de los campos de la Tabla Twitter2.

Figura 11. Descripción de los campos de la Tabla Volcán.

Figura 12. Descripción de los campos de la Tabla Terremoto.

Figura 13. Descripción de los campos de la Tabla Fuego.

Figura 14. Descripción de los campos de la Tabla Clima.

3.2. Consideraciones de diseño

3.2.1. Determinación de la ventana máxima de extracción de Tweets

Se conoce que la librería Tweepy y API de Twitter posee una limitación al momento de realizar las consultas, es decir para extraer información de la cuenta de usuario como también del Time line, entre otras acciones que se permiten realizar. Para lograr obtener este valor se realizaron varias pruebas para determinar el número de consultas máximo en función de un período entre consultas que se pueden realizar en un tiempo de 60 minutos, los resultados se detallan en la tabla 15.

Tabla 15.
Determinación del límite de consultas.

Tiempo entre consultas (min)	Número de consultas realizadas	Duración de la prueba (min)
1	15	15
5	13	60
10	7	60
15	5	60
20	4	60

Como resultado de la primera prueba, con tiempo entre consultas de 1 minuto se pudo determinar que el número máximo de consultas que se pueden realizar en 60 minutos. Después de este tiempo el API de Twitter presentó un error referente al límite

de consultas permitido y generó un bloqueo de otros 60 minutos en los cuales no se permite la realización de ningún tipo de consulta.

3.2.2. Determinación del período de extracción de Tweets

Para la determinación de la frecuencia de extracción de Tweets se tomaron en cuenta varios parámetros que son: el tiempo entre muestras para no afectar la ventana de operación de la librería y API, descrito en el escenario anterior; los días de prueba, esto en función de cuales son los días que se genera más tráfico en la aplicación; la hora del día en la que se toman las muestras; y por último un parámetro ambiguo que es la presencia de desastres naturales considerados eventos. Todo esto influye para la determinación del período.

Para la realización de esta consideración se realizaron varias pruebas en función de un día en específico, una hora, la cantidad de Tweets capturados y la variación en relación al anterior.

La primera prueba se realizó para el día 08/12/2015 a las 10H00 AM, con un tiempo entre capturas de 5 minutos con cinco ejecuciones del programa en un período de tiempo de 20 minutos, los resultados se detallan en la tabla 16.

Tabla 16.
Variación de captura de Tweets con tiempo entre captura de 5 minutos.

# de captura	# de Tweets capturados	# de Tweets repetidos	Tweets con contenido para filtrar
1	17	0	0
2	19	2	0
3	23	4	0
4	28	5	2
5	24	4	0

De los resultados obtenidos se puede observar que para un período de 20 minutos en ese día y a la hora descrita se tiene un promedio de 3 Tweets repetidos por el total del tiempo de captura y un promedio de 0.4 Tweets que deben ser filtrados. Para el caso de un decrecimiento en la captura de Tweets se debe a la inclusión de contenido con caracteres "emoji" o URL los cuales son eliminados en el proceso de extracción del contenido.

La segunda prueba se realizó para el día 12/12/2015 a las 12H00 PM, con un tiempo entre capturas de 10 minutos con cinco ejecuciones del programa en un período de tiempo de 40 minutos, los resultados se detallan en la tabla 17.

Tabla 17.
Variación de captura de Tweets con tiempo entre captura de 10 minutos.

# de captura	# de Tweets capturados	# de Tweets repetidos	Tweets con contenido para filtrar
1	8	0	0
2	28	8	0
3	30	2	0
4	25	5	0
5	24	1	0

De los resultados obtenidos se puede observar que para un período de 40 minutos en ese día y a la hora descrita se tiene un promedio de 3,2 Tweets repetidos por el total del tiempo de captura y un promedio de 0 Tweets que deben ser filtrados, lo que indica que a esa hora y en el período de captura descrito previamente no se registraron eventos. Para el caso de un decrecimiento en la captura de Tweets se debe a la inclusión de contenido con caracteres "emoji" o URL los cuales son eliminados en el proceso de extracción del contenido.

La tercera prueba se realizó para el día 22/12/2015 a las 10H00 AM, con un tiempo entre capturas de 15 minutos con tres ejecuciones del programa en un período de tiempo de 30 minutos, los resultados se detallan en la tabla 18.

Tabla 18.
Variación de captura de Tweets con tiempo entre captura de 15 minutos.

# de captura	# de Tweets capturados	# de Tweets repetidos	Tweets con contenido para filtrar
1	19	0	0
2	23	4	0
3	30	7	3

De los resultados obtenidos se puede observar que para un período de 40 minutos en ese día y a la hora descrita se tiene un promedio de 3,6 Tweets repetidos por el total del tiempo de captura y un promedio de 1 Tweets que deben ser filtrados, lo que indica que a esa hora y en el período de captura descrito previamente se registraron eventos.

La cuarta prueba se realizó para el día 26/01/2016 a las 08H00 AM, con un tiempo entre capturas de 20 minutos con 6 ejecuciones del programa en un período de tiempo de 100 minutos, los resultados se detallan en la tabla 19.

Tabla 19.
Variación de captura de Tweets con tiempo entre captura de 20 minutos.

# de captura	# de Tweets capturados	# de Tweets repetidos	Tweets con contenido para filtrar
1	20	0	0
2	20	0	0
3	20	0	0
4	19	1	1
5	19	0	1

6	19	0	0

De los resultados obtenidos se puede observar que para un período de 100 minutos en ese día y a la hora descrita se tiene un promedio de 0.16 Tweets repetidos por el total del tiempo de captura y un promedio de 0.33 Tweets que deben ser filtrados. Para el caso de un decrecimiento en la captura de Tweets se debe a la inclusión de contenido con caracteres “emoji” o URL los cuales son eliminados en el proceso de extracción del contenido.

3.3. Aplicación Web

Para la aplicación web se maneja un esquema de cinco páginas independientes donde existe un home o Index, en la cual se desarrolla de una forma gráfica la evolución de la base de datos y una página para cada categoría donde se despliegan los últimos Tweets registrados con la posibilidad de cargar todo el contenido de la categoría, al igual que un botón que permite regresar al home de la aplicación.

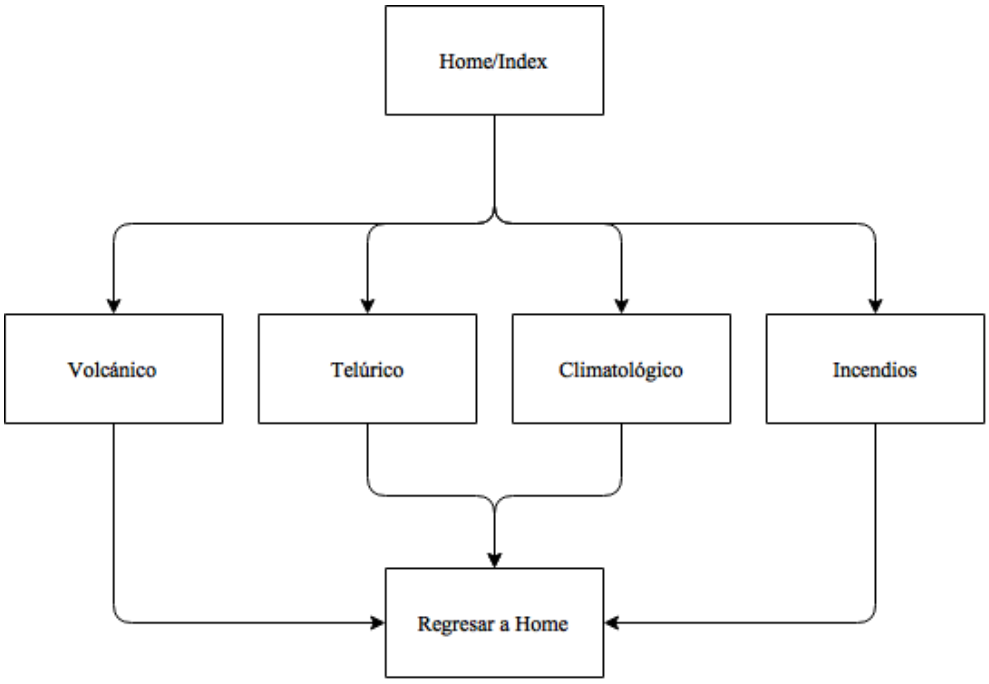


Figura 15. Esquema de navegación de AW.

Se decidió que todo el trabajo que no ve el usuario final se lo realice en Python, como se explicó en las secciones anteriores, y la parte visual para una mejor fluidez

como ya se mencionó en el Capítulo anterior, se implemente en HTML, PHP y algunos complementos escritos en JavaScript, logrando así una mayor velocidad y mejor manejo de los resultados el momento de subirlos a la aplicación, todo esto gracias a las ventajas que estos lenguajes presentan en el diseño Web sobre Python.

La conexión con la base de datos se realiza con el uso del lenguaje de programación PHP, donde se requieren parámetros previamente declarados como son: la dirección IP de la base de datos, el nombre de la misma, el puerto de acceso, el usuario y la contraseña de acceso del mismo. En base a un estilo declarado a objetos, son los parámetros requeridos por la clase "mysqli" para realizar la conexión con la base de datos. De ser estos correctos se efectua la conexión, caso contrario de despliega un mensaje de error en la conexión. Proceso descrito en la figura a continuación.

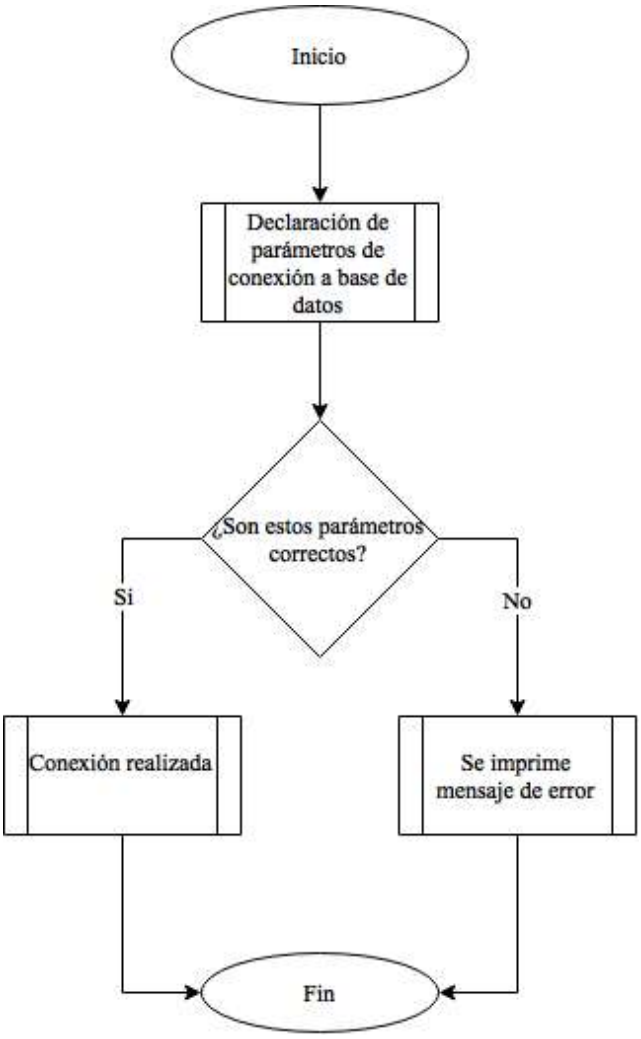


Figura 16. Proceso de conexión a base de datos.

Por cada categoría se realiza un “query” a la tabla que corresponde, en donde se selecciona en orden descendente los Tweets registrados para ser plasmados de una forma ordenada y bajo el criterio: fecha, autor del Tweet y el texto del Tweet tomando la representación gráfica de una tabla.

Para la realización de las gráficas se utiliza la herramienta Google Chart Tools, escrita en JavaScript, por la que se debe declarar una conexión en línea a la misma para acceder a las librerías que posee. Y mediante el uso de la función de dibujo de gráficas “drawChart” se realiza un arreglo matricial con el nombre de las categorías para la primera columna y para la segunda columna, con un código PHP se imprimen los valores referentes al total registros de cada categoría, esto se realiza mediante un “query” específico para consultar el último valor. Para la gráfica que relaciona el total de Tweets capturados con el número de DN registrados en total, se suman los últimos valores de cada categoría para determinar un resultado global.

Para esto se dividieron las acciones que se puede realizar dentro de la aplicación y los actores que participan con los respectivos casos que se pueden generar, como se detalla en la figura 17. El sistema siempre proporciona un menú que contiene las categorías de Desastres Naturales, el usuario puede seleccionar cualquiera indistintamente, y al realizar esta acción, se le proporciona/despliega la información sobre los sucesos reportados de la categoría seleccionada.

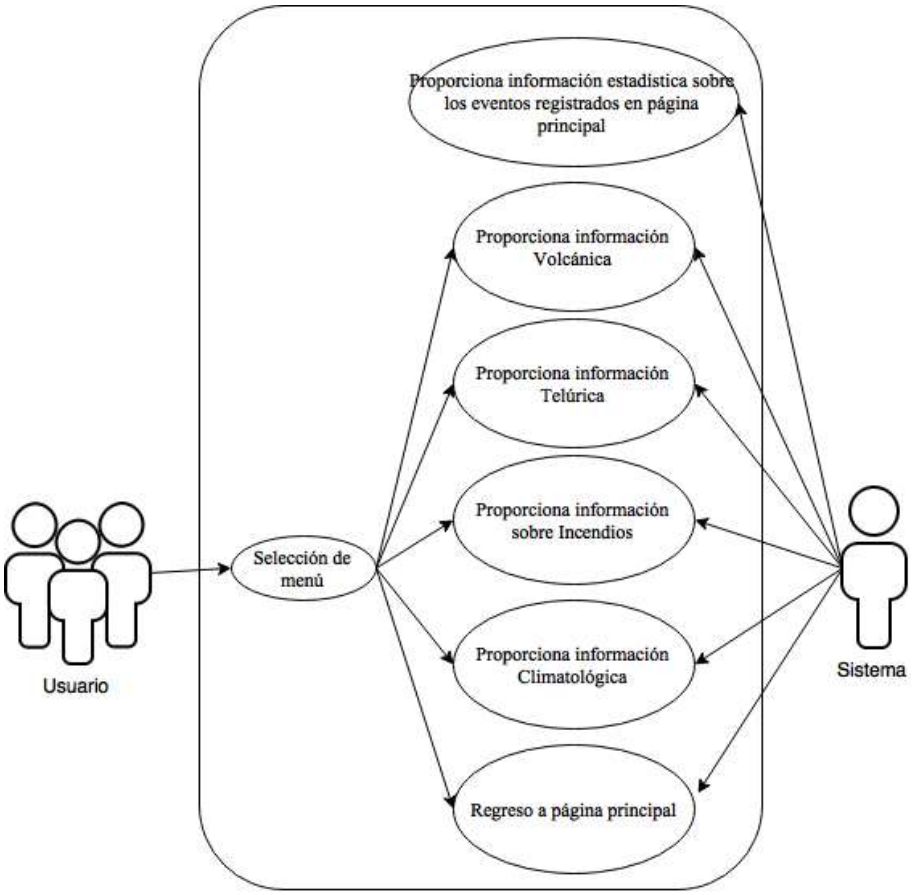


Figura 17. Diagrama de casos para la aplicación web.

Se divide el cuerpo de cada página en tres partes como se muestra en la figura 18, la cabecera contiene los títulos y los menús accesibles por el usuario; que al seleccionar uno de estos direcciona a otra página con la misma estructura pero que con la inserción de códigos PHP se accede a las bases de datos para obtener la información y desplegarla en el contenido, descrito previamente. La tercera sección es el pie de página, contiene dos botones: uno para Twittear una noticia relevante por parte del usuario y el otro para cargar el contenido de la base de datos correspondiente a la categoría seleccionada.

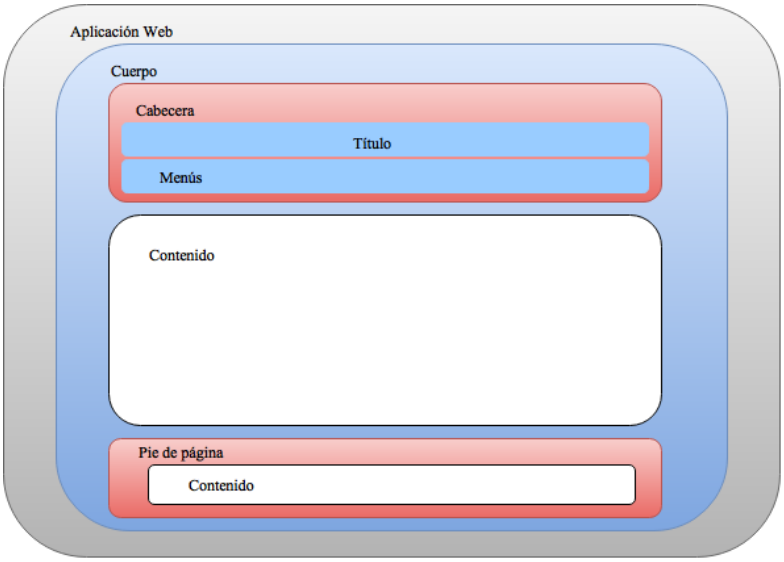


Figura 18. Esquema de página web de la aplicación.

3.4. Diseño final de la aplicación

Siguiendo el modelo planteado para la parte web en la sección anterior, la página principal de la aplicación o índex se ilustra en la figura a continuación:

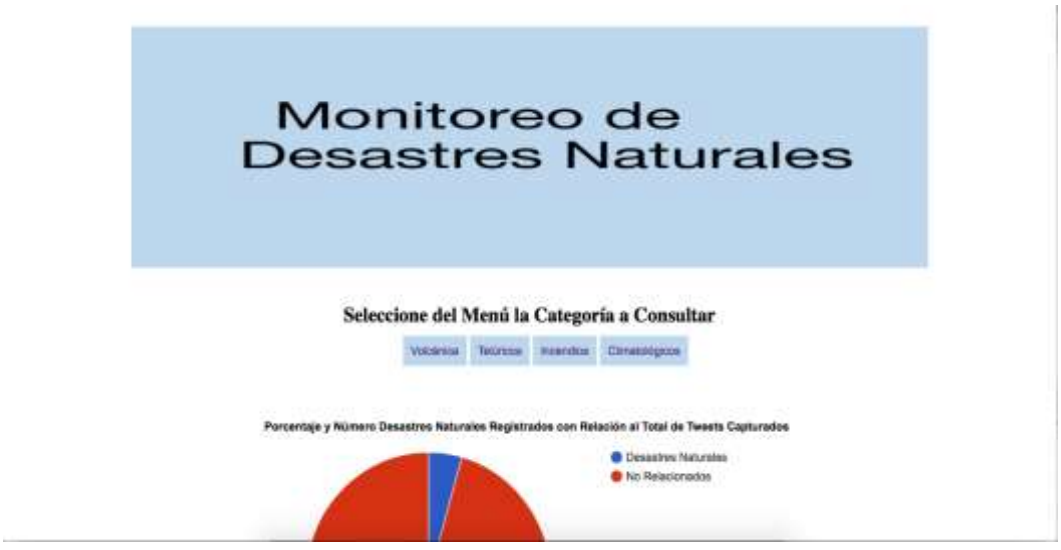


Figura 19. Índex o Home de la aplicación web.

A continuación se ilustra el contenido por cada categoría, comenzando con la categoría Volcán, Telúricos, Incendios y Climatológicos en las figuras 20, 21, 22 y 23 respectivamente.

Figura 20. Visualización del contenido de categoría Volcán.

Figura 21. Visualización del contenido de categoría Terremoto.

Figura 22. Visualización del contenido de categoría Incendios.

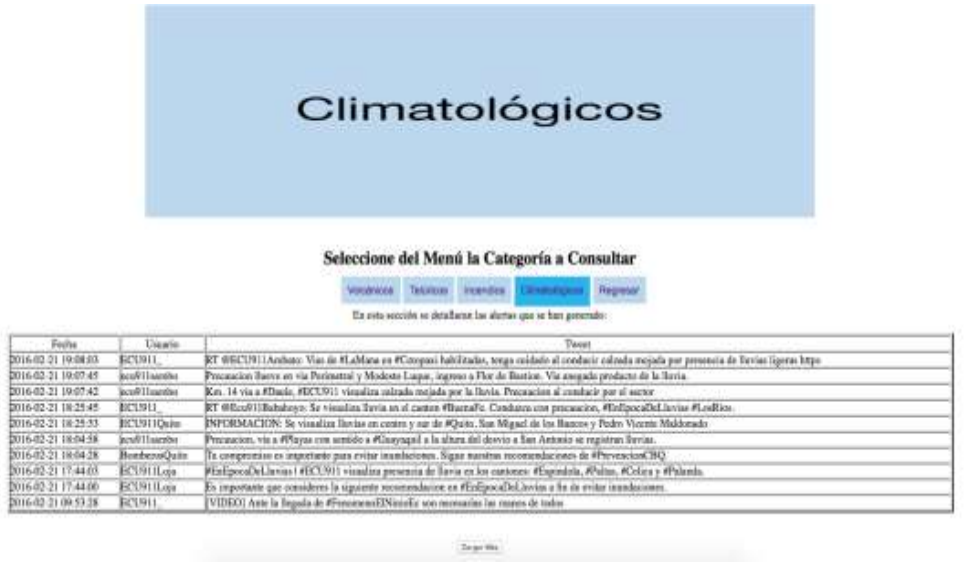


Figura 23. Visualización del contenido de categoría Clima.

CAPÍTULO 4

4.1. Escenarios, pruebas y resultados

Dentro de este Capítulo, se plantea la creación de diversos escenarios para la ejecución de la aplicación, al igual que diversas pruebas para comprobar su funcionamiento y por último un análisis de los resultados en función de los parámetros planteados y datos obtenidos.

4.1.1. Escenarios de prueba

4.1.1.1. *Escenario 1: Eficiencia del proceso de filtrado*

Se diseñó un algoritmo para el proceso de filtrado en función de un conjunto de palabras en relación a la categoría, como se lo describió en el Capítulo anterior, y para determinar la eficiencia de este se toma en cuenta el número de Tweets en una categoría con contenido erróneo sobre el número total de Tweets de la tabla a la que pertenece, todo esto dentro de la base de datos hasta el día 21 de Febrero del 2016. Obteniendo los resultados que se detallan en la tabla a continuación.

Tabla 20.
Relación de Tweets con contenido erróneo dentro de una categoría.

Categoría	Tweets con contenido erróneo	Tweets totales filtrados	Error (%)	Eficiencia (%)
Volcán	13	104	12.4	87.5
Terremoto	0	50	0	100
Incendios	12	95	12.64	87.36
Clima	2	255	0.78	99.22

Los principales problemas para el algoritmo de filtrado son palabras que forman un contexto diferente al que se desea filtrar pero poseen una o varias de las declaradas

para el proceso, más específicamente en la categoría Incendios la mayor cantidad de contenido erróneo se generó cuando se filtró sentencias con las palabras “armas de fuego” que si bien poseen una de las palabras seleccionadas para el proceso, su contexto no es el adecuado. Para la categoría Volcán, el principal contenido erróneo registrado al realizar el proceso de filtrado de información se genera con los nombres de las provincias, esto por estar relacionados con los nombres de los volcanes.

Para determinar la eficiencia del algoritmo a nivel general se promediaron los resultados obtenidos tal como se detalla en la tabla 21.

Tabla 21.
Eficiencia final del algoritmo de filtrado.

Error (%)	**6.45**
Eficiencia (%)	**93.55**

El contenido erróneo puede ser eliminado de las tablas mediante un script de búsqueda del mismo.

4.1.1.2. *Escenario 2: Visualización de la aplicación*

Actualmente acceder a una página web puede realizarse desde casi cualquier dispositivo electrónico, es decir: computadora, televisiones inteligentes, Tablet, celulares inteligentes, entre otros. Cada uno con su respetivo tamaño de pantalla.

Para realizar las pruebas de visualización se seleccionó las herramientas más comunes para acceso a internet, las cuales son: computador, Tablet y el celular inteligente. El tamaño de las pantallas, el sistema operativo y la versión del navegador usados para la realización de este escenario se describen en la tabla a continuación.

Tabla 22.
Tamaños de pantalla de los dispositivos usados.

Dispositivo	Tamaño de pantalla (pulgadas)	Sistema Operativo	Navegador de Internet	Versión Navegador
Computador	13	Mac OS 10.11.3	Safari	9.0.3
Tablet	7.9	IOS 9.2	Safari	9.2
Celular	4.5	Android 5.1	Chrome	48.0.2564.95

La visualización en cada uno de los dispositivos se detalla en las figuras 24, 25 y 26 donde se observa la pantalla principal (PP) de la aplicación web (AW) con los menús que llevan a cada categoría.

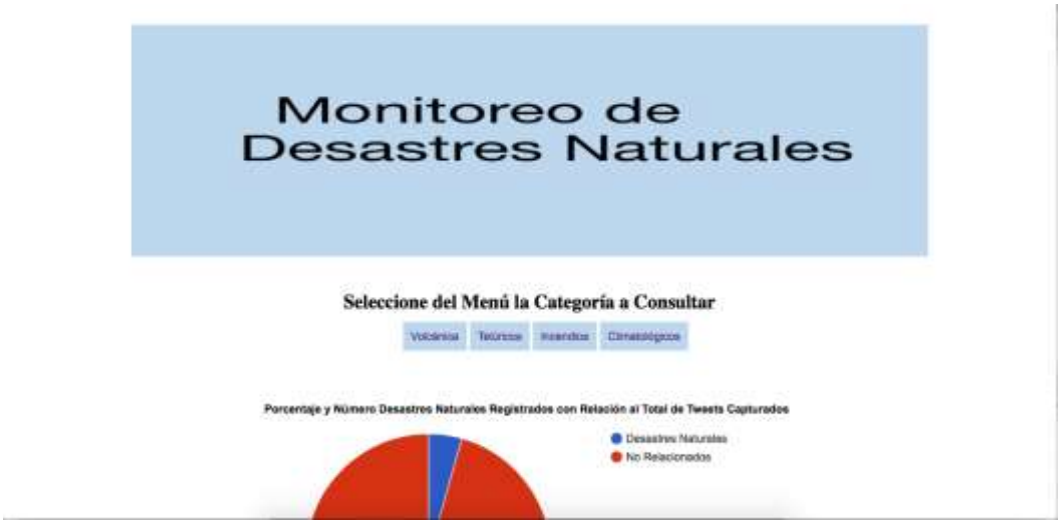


Figura 24. PP de la AW ejecutada en una computadoras de escritorio.

El contenido se adapta al tamaño de la pantalla y permite su correcta visualización al igual que a otros beneficios que un computador de escritorio presta como: el uso del "mouse" para poder desplazarse por la página al igual que la aplicación del zoom.

Figura 25. PP de la AW ejecutada en una Tablet.

El contenido se adapta completamente a la pantalla de la Tablet permitiendo una visualización plena de toda la página web.

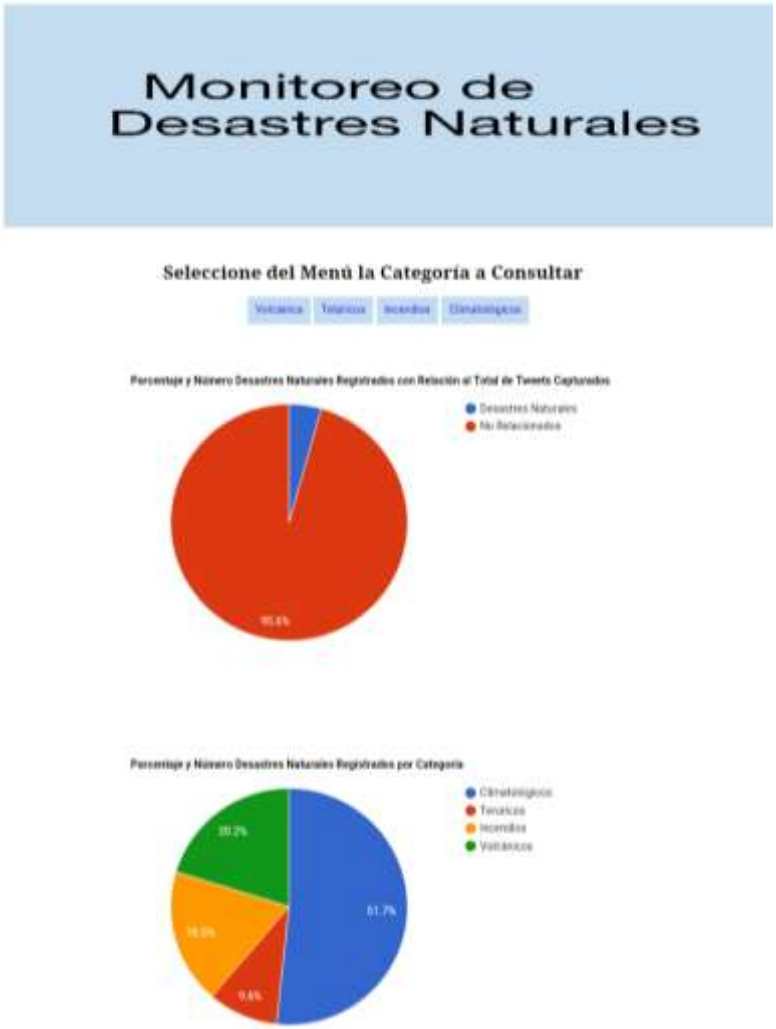


Figura 26. PP de la AW ejecutada en un celular inteligente.

Para el teléfono inteligente el contenido de la página desarrollado en HTML se adapta al tamaño de la pantalla, en esta se puede visualizar los aplicativos de complementos escritos en otro lenguaje de programación, como es el caso del menú realizado en JavaScript, no se ajusta completamente a la nueva dimensión y mantiene la diseñada.

Para el análisis de visualización de contenido se seleccionó la categoría con más registros almacenados en la base de datos, buscando un límite máximo de carga de contenido para cada uno de los dispositivos, como se muestra en las figuras 27, 28 y 29 respectivamente.

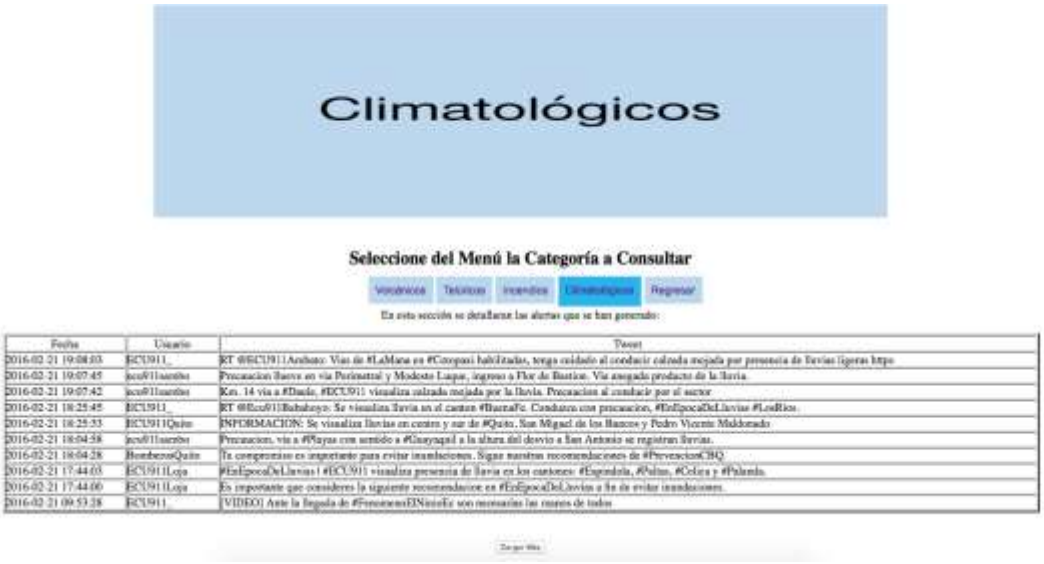


Figura 27. Categoría clima en un computador de escritorio.

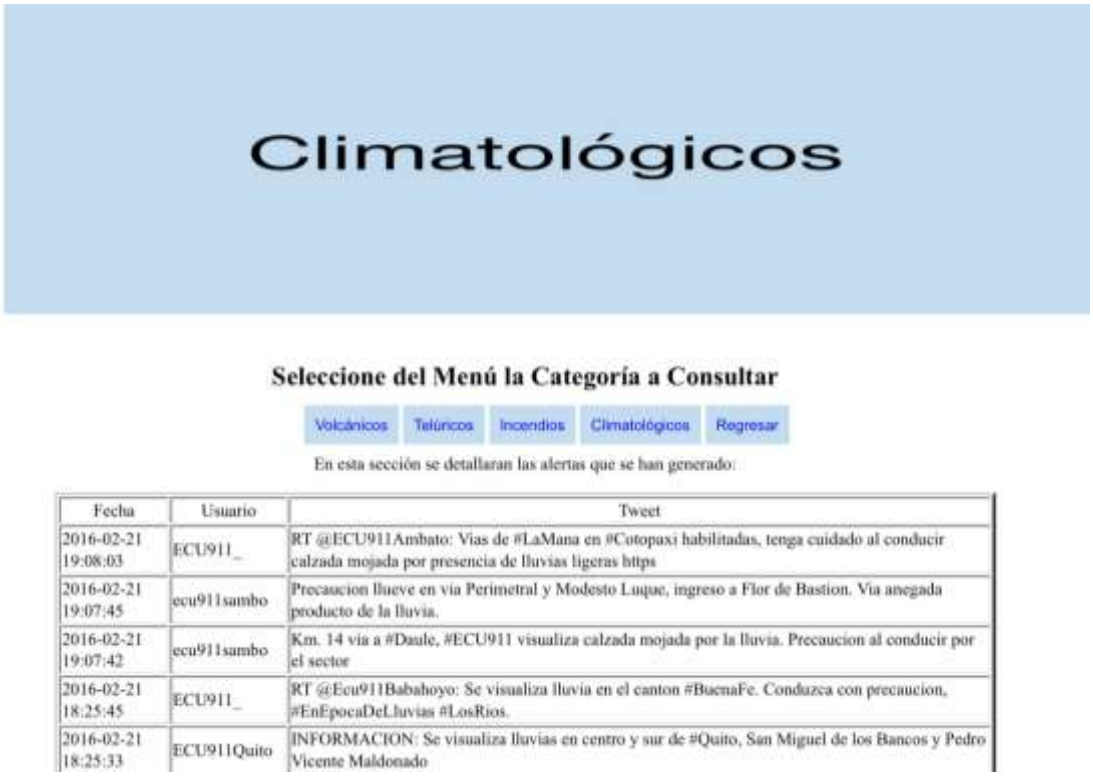


Figura 28. Categoría clima en un explorador de Internet en una Tablet.

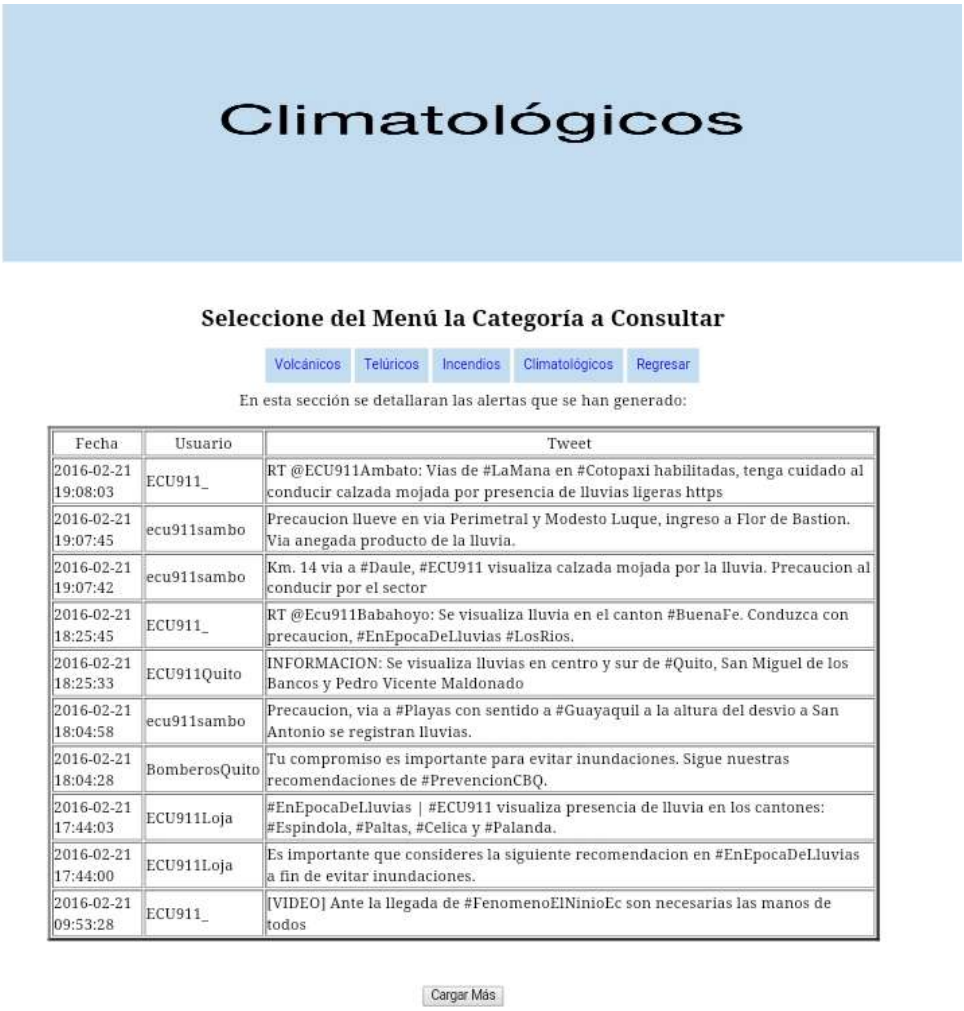


Figura 29. Categoría clima en un explorador de Internet en celular inteligente.

Se observó que el contenido de la página web, al momento de seleccionar las consultas se adapta mejor al navegador de Internet, que en comparación con los dispositivos móviles, estos por el tamaño de su pantalla se ven afectados en la longitud horizontal del texto que puede presentar, al igual que la necesidad de requerir un mayor espacio vertical para realizar esta misma acción. Al incrementar el número de consultas que se pueden visualizar se necesita un desplazamiento por la página, acciones que pueden incomodar y también se pueden generar reportes de acontecimientos muy antiguos.

4.1.1.3. *Escenario 3: Pruebas del servidor*

Para este escenario se plantearon 3 pruebas diferentes para montar el servidor y probar el correcto funcionamiento de la aplicación en los diferentes dispositivos y en la red.

La primera prueba consiste en montar el servidor en la misma máquina donde se desarrolló el proyecto para una red interna. Las pruebas se pueden realizar desde el navegador accediendo a la dirección local de la computadora (localhost o 127.0.0.1) y desde otros dispositivos en la misma red con la dirección IP asignada por el Router a la misma. Esta prueba solo se puede realizar dentro de una red local.

La segunda prueba consiste en montar toda la aplicación en servidores externos gratuitos, para esto se debe tomar en cuenta un host que permita el acceso a una base de datos MySQL, que pueda procesar Python para la captura de Tweets y PHP para el desarrollo web. Esta prueba permitiría la realización de la aplicación desde cualquier dispositivo en cualquier parte del mundo.

El proceso consistió en la búsqueda de un servidor/host que permita la integración de los lenguajes mencionados, los resultados se detallan en la tabla 23.

Tabla 23.
Comparativa entre los servicios que ofrecen Host para páginas Web.

Host	PHP	MySQL	Python	Gratuito
Hostinger	Si	Si	No	Parcial
260mb	Si	Si	No	Limitado a 260 Mb
Wordpress.org	Si	Si	No	Más desarrollado para Blogs
Wordpress.com	Si	Si	No	$99
Joomla	Si	Si	No	Gratuito

Hostinger proporciona una mayor cantidad de ventajas de las necesitadas, pero la creación de dominios para el Ecuador no es gratuita y solo permite un subdominio como página principal.

De las investigadas, Python no es soportado para montarlo como parte del motor de búsqueda y descarga de Tweets, esto debe realizarse aparte. Con este inconveniente solucionado se puede trabajar en cualquiera de los host mencionados, los más Wordpress.org y Joomla, pero esto implica la compra de un dominio propio para poder montar el servidor y permitir el acceso a la base de datos generada por el script en Python.

La tercera prueba consiste en continuar usando la computadora donde se desarrolló la aplicación como el servidor pero requiere abrir el puerto de servidor web del equipo (80) usado por el proveedor de Internet para domicilios.

El proveedor de servicio de internet utilizado para esta prueba proporciona una IP dinámica a sus clientes, esto ya impone un primer inconveniente el momento de acceder a la aplicación web ya que se debería conocer cada cuanto tiempo cambia la dirección el proveedor al cliente, esto se realiza una vez por semana y se asigna generalmente la misma dirección IP pero si existen casos donde se renuevan las direcciones, información que fue negada a entregarse, por lo cual no se puede especificar el día en el que se realiza esta acción. El segundo y mayor inconveniente fue que al haber contratado un plan residencial la libre configuración del equipo no es permitida la manipulación de los mismos por parte del usuario, realizar esta acción puede generar costos y multas por mal uso del servicio ya que el puerto 80 se encuentra bloqueado y esto se especifica en el contrato.

4.1.1.4. *Escenario 4: Comportamiento Social*

Durante el proceso de elaboración del proyecto se registraron varios hechos no relacionados con Desastres Naturales, dentro de los más importantes, se puede hablar de dos en concreto: uno político a nivel nacional y otro de farándula internacional; los

cuales fueron registrados en un escenario de prueba para medir la cantidad de Tweets generados durante el período de análisis realizado al igual que la cantidad de Re Tweets que los mismos registran para compararlos con los obtenidos bajo el mismo criterio pero bajo el proceso de filtrados por categoría, es decir, los que contienen información sobre Desastres Naturales, en la tabla a continuación de detalla las características de los eventos para los días 12, 15 y 16 de Febrero del 2016.

Tabla 24.
Tipo de evento registrado para los días de observación.

Categoría	**Tipo de evento registrado**
Política	Manifestación Colegio Montufar
Social	Grammy 2016
Volcánico	Días normales de observación
Telúrico	Días normales de observación
Incendio	Pocos eventos registrados
Climatológico	Días normales de observación

Los resultados obtenidos para las nuevas categorías política y social se describen en las figuras 30 y 31 respectivamente, durante este período se puedo determinar un gran incremento de Tweets lo que implica un gran seguimiento de los usuarios o fuentes de reporte de información para la realización de una completa cobertura para estos tipos de eventos que son completamente diferentes y en distintos lugares del mundo.

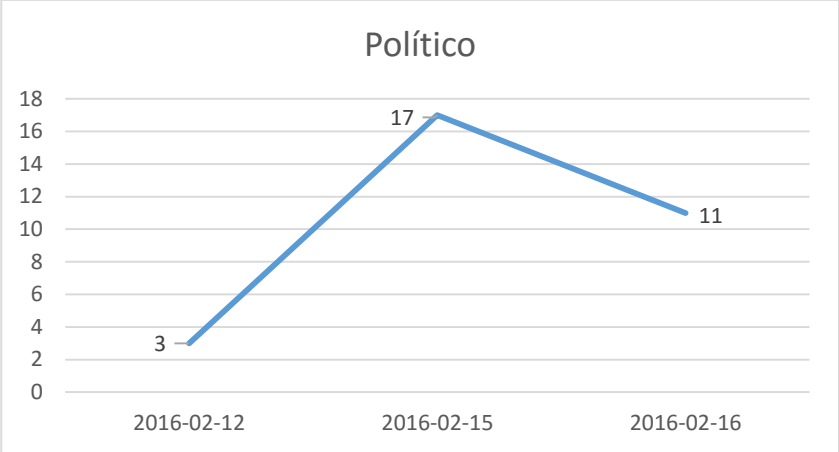


Figura 30. Tendencia registrada sobre un tema político.

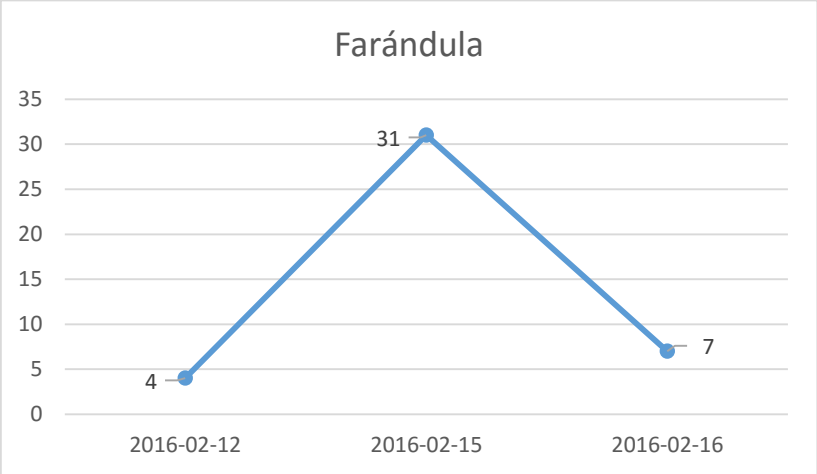
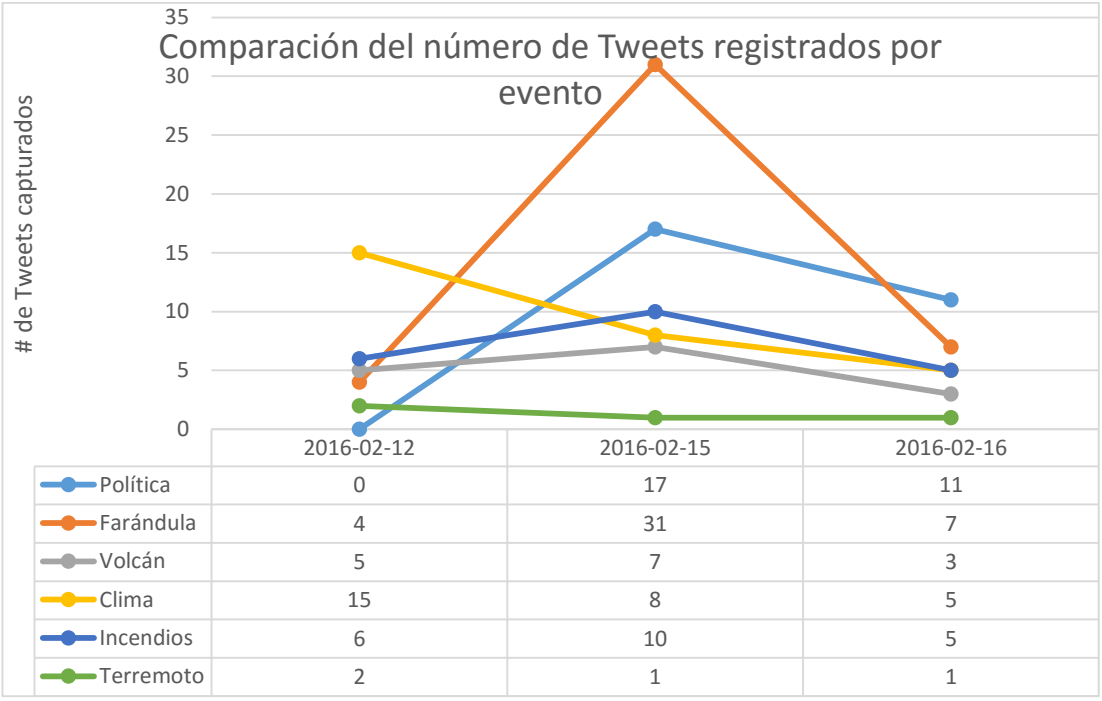


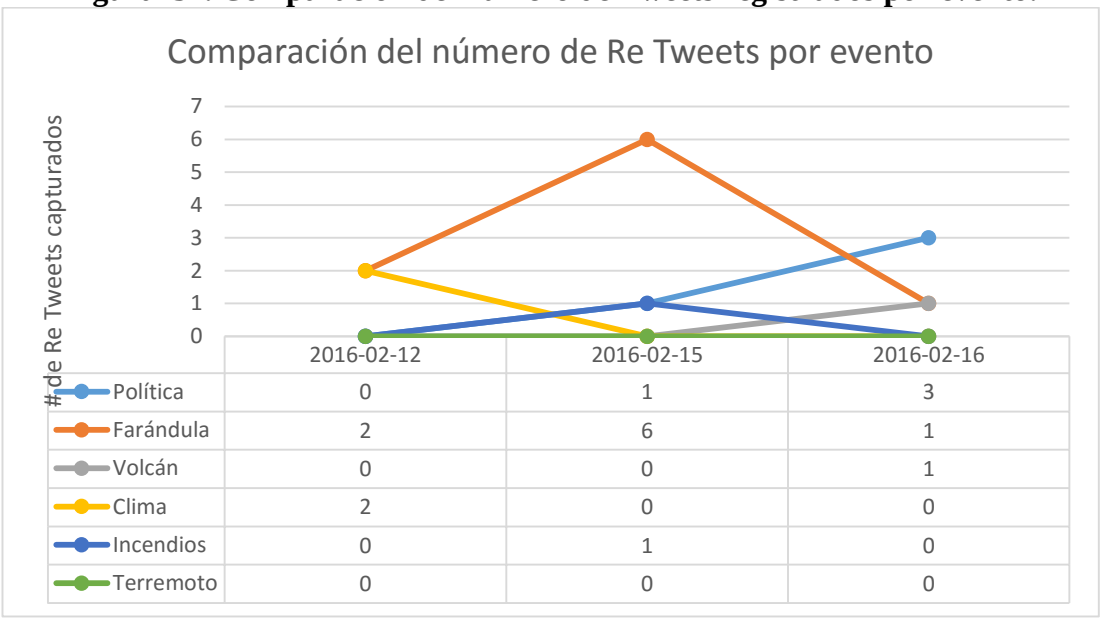
Figura 31. Tendencia registrada sobre un tema social.

Se puede observar el incremento de Tweets el día 15 de Febrero del 2016, fecha en la que se ocurrieron los eventos. Estos resultados fueron comparados con los Tweets de cada categoría y esto de describe en la figura 32, al igual que la cantidad de Re Tweets que generaron los eventos, describiéndose este caso en la figura 33.



	2016-02-12	2016-02-15	2016-02-16
Política	0	17	11
Farándula	4	31	7
Volcán	5	7	3
Clima	15	8	5
Incendios	6	10	5
Terremoto	2	1	1

Figura 32. Comparación del número de Tweets registrados por evento.



	2016-02-12	2016-02-15	2016-02-16
Política	0	1	3
Farándula	2	6	1
Volcán	0	0	1
Clima	2	0	0
Incendios	0	1	0
Terremoto	0	0	0

Figura 33. Comparación del número de Re Tweets por evento.

De las figuras se puede apreciar que los usuarios prefieren dar más seguimiento a eventos sociales que a la cobertura de Desastres Naturales.

4.2. Resultados estadísticos

En esta sección se analiza estadísticamente los datos generados en la aplicación de una forma gráfica mientras que los datos detallados se encuentran en el Anexo 5; es decir, en esta parte del documento se comenta la cantidad de datos obtenidos en un período de tiempo y cuales de estos datos corresponden a desastres naturales que debieron ser filtrados y clasificados dentro de las categorías mostrando la evolución de la base de datos.

4.2.1. Tabla Twitter

Se resumió el número de Tweets generados hasta la fecha 21 de Febrero del 2016 en la Tabla Twitter, donde se detalla cada día de captura con el número de Tweets que se registraron indistintamente y sin clasificar como se describe de una forma gráfica en la figura 34, dando como resultado un promedio de 322.36 Tweets capturados por día.

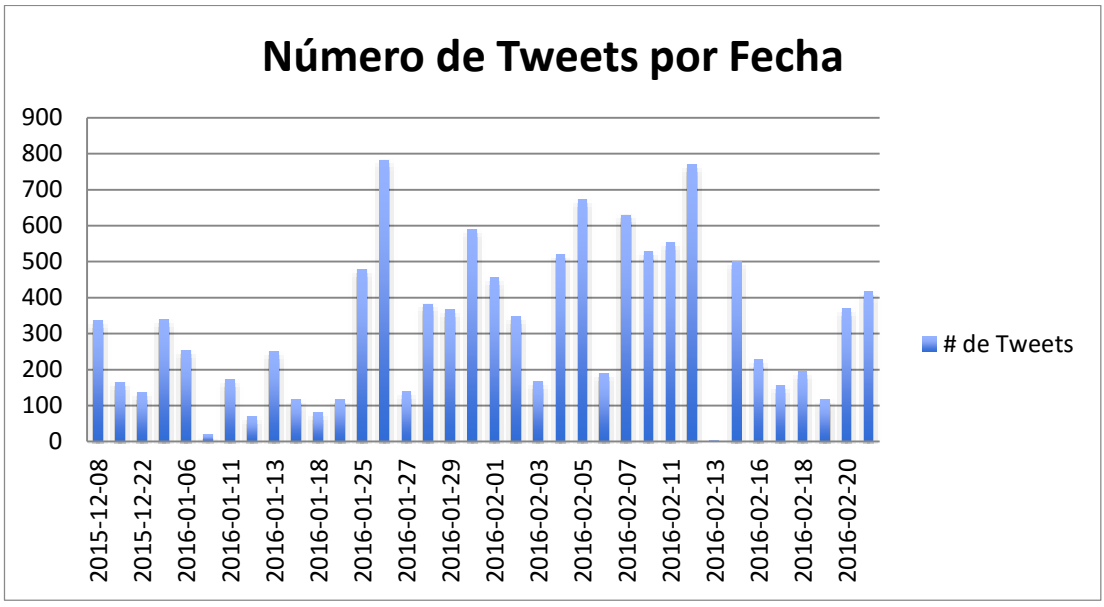


Figura 34. Detalle del número de Tweets capturados para la tabla Twitter2.

4.2.2. Tabla Volcán

Se resumió el número de Tweets generados hasta la fecha 21 de Febrero del 2016 en la Tabla Volcán, donde se detalla cada día de captura con el número de Tweets que se registraron dentro de esta categoría como se describe de una forma gráfica en la figura 35, dando como resultado un promedio de 3.71 Tweets relacionados a esta categoría siendo los principales Tweets los informes diarios de monitoreo de los volcanes.

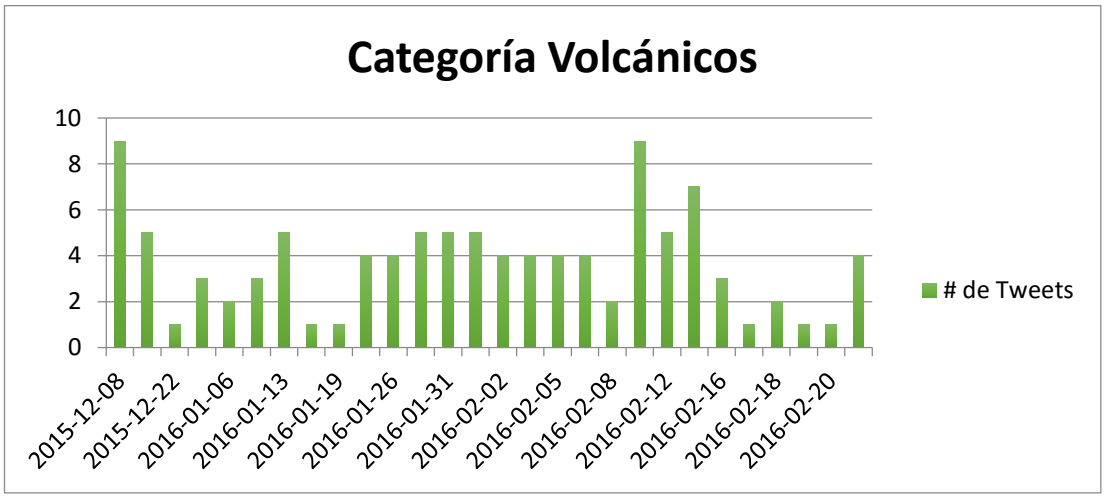


Figura 35. Detalle del número de Tweets capturados para la tabla Volcán.

4.2.3. Tabla Terremoto

Se resumió el número de Tweets generados hasta la fecha 21 de Febrero del 2016 en la Tabla Terremoto, donde se detalla cada día de captura con el número de Tweets que se registraron dentro de esta categoría, como se describe de una forma gráfica en la figura 36, dando como resultado un promedio de 3.07 Tweets relacionados a esta categoría, el bajo número de capturas en esta categoría indica una reducción en la actividad de estos desastres naturales en este período de tiempo.

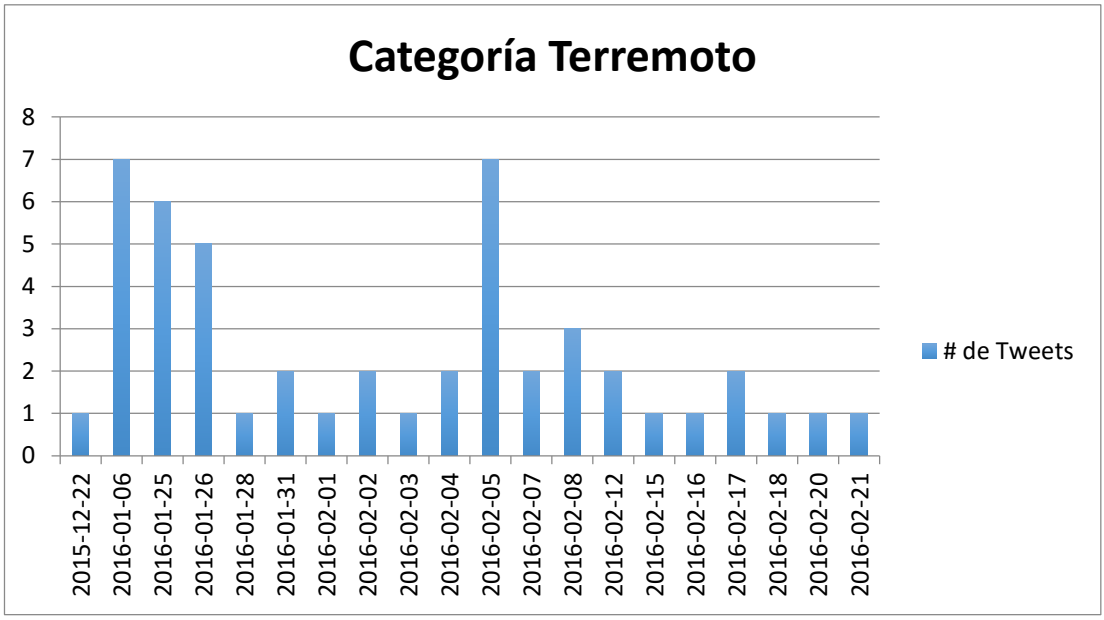


Figura 36. Detalle del número de Tweets capturados para la tabla Terremoto.

4.2.4. Tabla Clima

Se resumió el número de Tweets generados hasta la fecha 21 de Febrero del 2016 en la Tabla Clima, donde se detalla cada día de captura con el número de Tweets que se registraron dentro de esta categoría, como se describe de una forma gráfica en la figura 37, dando como resultado un promedio de 7,9 Tweets relacionados a esta categoría, se puede observar que existe un crecimiento durante los días lluviosos (25 y 26 de Enero del 2016) en el país sin ser estos ocasionados por el Fenómeno de El Niño, y principalmente en la región costa.

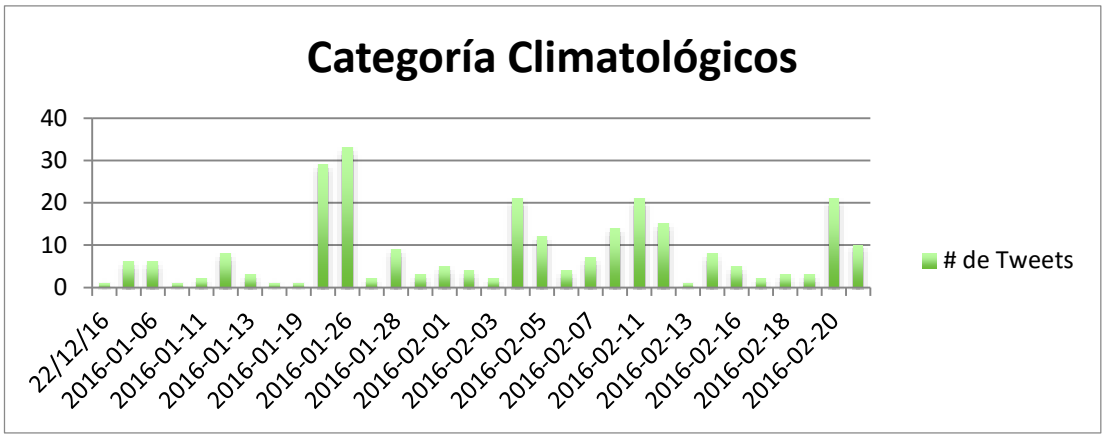


Figura 37. Detalle del número de Tweets capturados para la tabla Clima.

4.2.5. Tabla Fuego

Se resumió el número de Tweets generados hasta la fecha 21 de Febrero del 2016 actual en la Tabla Fuego, donde se detalla cada día de captura con el número de Tweets que se registraron dentro de esta categoría, como se describe de una forma gráfica en la figura 38, dando como resultado un promedio de 4.92 Tweets relacionados a esta categoría, varía en función del clima, siendo escaso en días lluviosos al igual que varía por los efectos solares como también por las acciones causadas por el hombre.

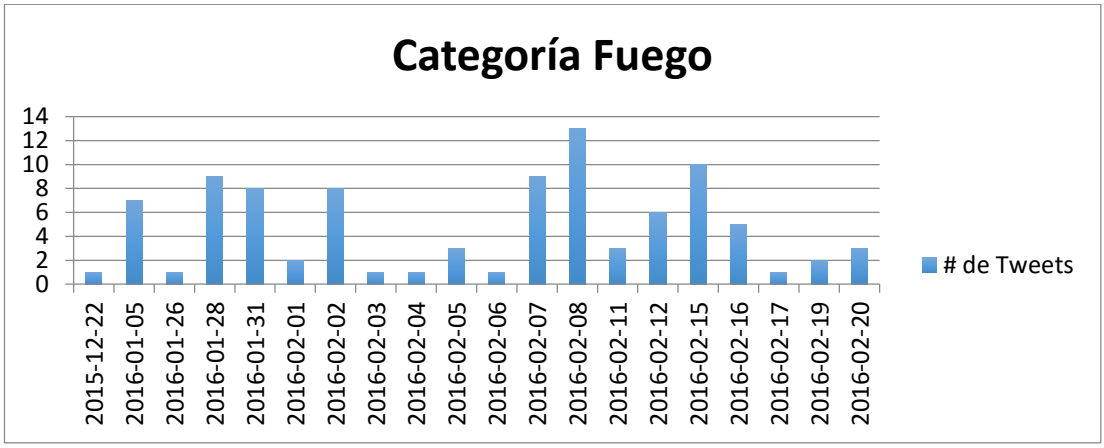


Figura 38. Detalle del número de Tweets capturados para la tabla Fuego.

4.2.6. Resumen General

Para poder comparar los resultados estadísticos se realizó un análisis por categoría en función del total de Tweets obtenidos y su porcentaje total, dando un resultado como el descrito en la tabla a continuación y de una forma gráfica en la figura 39.

Tabla 25.
Resumen actual de los datos generados en la aplicación.

Categoría	# de Tweets	Porcentaje
Twitter	11605	100
Volcán	104	0,896165446
Terremotos	49	0,422231797
Clima	263	2,266264541
Fuego	94	0,809995692
No relacionados	11095	95,60534252

Figura 39. Relación de Tweets No relacionados y Tweets DN.

Para determinar el número de Desastres Naturales que se registraron en el período de tiempo de captura se sumó todas las categorías al igual que su porcentaje como se describe en la tabla 26, de donde se puede decir que los Tweets relacionados con desastres naturales no sobrepasan al 5 %, todo esto en función a un usuario promedio con seguimiento a cuentas de noticias, organismos de socorro, canales de televisión, periódicos, entretenimiento, deportes, entre otros.

Tabla 26.
Número de Tweets sobre DN y su porcentaje con relación al total.

Total DN	510
Porcentaje DN	4,394657475

Dentro de las categorías los porcentajes que representa cada evento dentro del total de desastres naturales se describe en la tabla 27 y en la figura 40.

Tabla 27.
Porcentaje representativos de cada categoría.

Categoría	Porcentaje
Volcán	20,39215686
Terremoto	9,607843137
Clima	51,56862745
Fuego	18,43137255

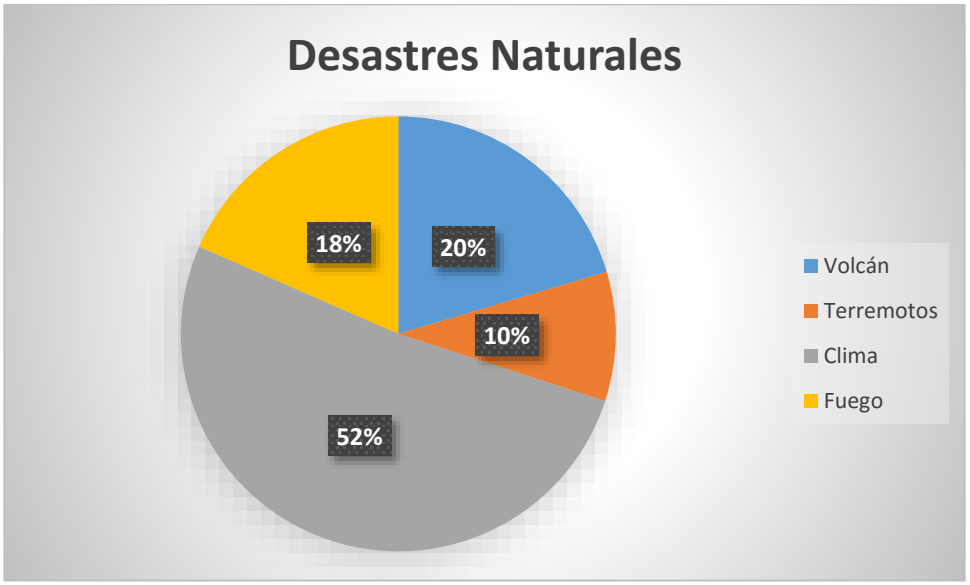


Figura 40. Porcentajes de las categorías de los DN registrados.

De donde se puede deducir que los eventos climatológicos representan aproximadamente el 50% de los sucesos registrados, siendo mayoritariamente las inundaciones causadas por las lluvias.

CAPÍTULO 5

5.1. Conclusiones

- Para el cálculo del tamaño de la ventana de operación de la librería Tweepy y el API de Twitter, se determinó que el valor de la misma es de 15 consultas en el período de 60 minutos o una hora, debido a que al superar el número de consultas se genera un bloqueo en el API prohibiendo realizar alguna acción de conexión o consulta durante un período de 60 minutos.
- Se determinó que el período entre extracción de Tweets debe ser de 20 minutos, debido al bajo promedio de Tweets repetidos o nulos que se registran y la baja pérdida de información que en este intervalo se generó.
- Se determinó que la eficiencia del algoritmo de filtrado de Tweets supera el 93% y se considera adecuado para la realización del proyecto, todo esto debido a que las fallas generadas son errores de contexto del contenido de una frase y de uso que se les da a las palabras.
- De las pruebas realizadas para determinar el límite del contenido adecuado de forma vertical (número de Tweets) para mostrar en pantalla, se delimitó a 10 ítems por categoría y en orden descendente, para facilitar la visualización en todos los dispositivos sin necesidad de crear versiones exclusivas para cada uno de estos (Computador, Tablet o Celular móvil).
- Las pruebas de funcionamiento del servidor de la aplicación solo se pudieron realizar a nivel de una red local, debido a las limitaciones presentadas por los host de Internet al no soportar la ejecución de Python y el proveedor del servicio en la negativa de la apertura del puerto 80.
- Al analizar los resultados de las pruebas realizadas sobre el comportamiento social, se puede afirmar que los usuarios de la red social Twitter dan una mayor cobertura a eventos de farándula que a eventos relacionados a Desastres Naturales o Política, en base a la cantidad de Re Tweets que los eventos registraron.
- Para realizar el registro de varios usuarios y personalización de la aplicación se requiere que los mismos posean una cuenta de desarrollador en Twitter, esto

se debe a las limitaciones de la librería Tweepy, que como parámetros de conexión requiere de cuatro permisos de autorización que son obtenidos como desarrollador de aplicaciones para la red social.

- El API de Twitter permite extraer información sobre un Tweet, como es: la fecha de creación, el usuario que generó el Tweet y la ubicación geográfica donde se realiza el post. Esta última información debe ser activada por los usuarios en la configuración de la cuenta, debido a esto no se pudo completar la realización un mapa de eventos en tiempo real sino uno actualizado manualmente.
- Al analizar la base de datos y los resultados obtenidos se pudo determinar que los organismos de seguridad y control como también las instituciones periodísticas son los principales agentes de brindan cobertura y seguimiento sobre la información de Desastres Naturales y no los usuarios como se pensó, debido a que los mismos usan más la red social como medio informativo.

5.2. Recomendaciones

- Debido a la cantidad de parámetros (fechas, usuarios, Tweets) que se analizaron, se recomienda la exportación de los datos a un documento de Excel que a un archivo CSV, para facilitar el manejo de los mismos.
- Se recomienda la utilización del lenguaje de programación Python para la realización de este tipo de proyectos debido a la gran cantidad de documentación y su fácil manejo.
- Se recomienda mantener actualizadas las librerías ya que continuamente estas incluyen mejoras, actualización de métodos, corrección de errores que pueden contribuir a simplificar el desarrollo de un proyecto.

5.3. Trabajos Futuros

- Se propone la realización de un estudio y análisis prolongado sobre el comportamiento social en su uso de la aplicación Twitter, como medio de información e interacción con los usuarios.
- NLTK es una herramienta completamente desarrollada para el idioma Inglés donde se permite analizar el contexto de una frase, texto u oración de una forma gramatical, pero para el Español estas capacidades no están implementadas en su totalidad, es por eso que se propone complementar a la herramienta para poder analizar el contexto del texto a analizar para mejorar el proceso de filtrado.
- Se propone la creación de aplicaciones móviles para los sistemas operativos predominantes en el mercado (IOS, Android), de una manera cliente/servidor, donde el teléfono móvil servirá solo de cliente y podrá visualizar los resultados en tiempo real; y que la obtención de datos se realice en un servidor dedicado.

REFERENCIAS

Ministerio Cordinador de Seguridad. (2014). *Valores/MIsión/Visión*. Obtenido de http://www.seguridad.gob.ec/el-ministerio/

Secretaría de Gestión de Riesgos . (2014). *Objetivos de la Secretaría de Gestión de Riesgos.*

Ministerio Cordinador de Seguridad. (2013). *Informe: Alertas, Incidentes y Despachos atendidos por el ECU911 Quito.*

Dirección Nacional de Defensa Civil. (2015). *Programa Sistema de Alerta Temprana Y Gestión de Riesgo Natural*. Obtenido de http://idbdocs.iadb.org/wsdocs/getdocument.aspx?docnum=561503

Secretaría de Gestión de Riesgos. (2010). *Implementación del sistema de alerta temprana en la cuenca del río Zarumilla.*

UNESCO. (2011). *Fortalecimiento del sistema Regional de Alerta Temprana ante Tsunami, preparativos en Chile, Colombia, Ecuador y Perú (2012-2013).*

Elguea, J. (1987). *Inteligencia Artificial y la sicología: Breve historia de la Inteligencia Artificial.* ITAM.

Hernández, M., & Gómez, J. (2013). Aplicaciones de Procesamiento de Lenguaje Natural. *Revista Politécnica*, 32.

Ferro, J. V. (2005). *Aplicaciones del Procesamiento del Lenguaje Natural en la Recuperación de Información en Español.* Universidade Da Coruña, Departamento de Computación. Universidade Da Coruña.

Alberich, M. (2007). *Procesamiento del Lenguaje Natural: Guía Introductoria.*

Otero, P. G., & González, M. G. (2000). *Técnicas de Procesamiento del Lenguaje Natural en la Recuperación de Información.* Centro de Investigación sobre Tecnoloxías da Lingua.

Rodríguez, S., & Carretero, J. (2010). *COES: Información General Y Distribución.*

Clarke, J., & Monstesinos, M. (2015). *Estudio Redes Sociales de IAB Spain.*

Gelbukh, A. (2010). *Procesamiento de Lenguaje Natural y sus Aplicaciones* . Komputer Sapiens . Komputer Sapiens .

Trabazos, O., Suárez, S., Bori, R., & Flo, O. (2014). Aplicación de tecnologías de Porcesamiento de Lenguaje Natural y tecnología semántica en Brand Rain Anpro21. *TALN*.

Glez, D. (2010). *Modelo para la integración de conocimiento biológico explícito en técnicas de clasificación aplicadas a datos procedentes de microarrays de ADN.* Universidad de Vigo. Universidad de Vigo.

Google-Inc. (2015). *Nuestra historia en profundidad*. Obtenido de https://www.google.com/about/company/history/?hl=es

Pérez, J. (1996). *Reconocimiento y generación integrada de la morfología del español: Una aplicación a la gestión de un diccionario de sinónimos y antónimos.* Univeridad de las Palmas de Gran Canaria, Facultad de Infromática.

Graña, J. (2002). *Técnicas de Análisis Sintáctico Robusto para la Etiquetación del Lenguaje Natural.* Universidad de La Coruña. Universidad de La Coruña.

Dubiau, L. (2013). *Procesamiento de Lenguaje Natural en Sistemas de Análisis de Sentimientos.* Universidad de Buenos Aires, Facultad de Ingeniería. Universidad de Buenos Aires.

Alcázar, S. (2013). *Diseño e implementación de un sistema para el análisis y categorización en Twitter mediante ténicas de clasificación automática de textos.* Universidad Carlos III de Madrid, Departamento Ingeniería Telemática. Universidad Carlos III de Madrid.

TensorFlow. (2015). *TensorFlow is an Open Source Software Library for Machine Intelligence*. Obtenido de https://www.tensorflow.org

University, S. (1983). *Natural Language Processing*. Obtenido de https://searchworks.stanford.edu/view/8196284

Román, J. V., García, R. C., & Rueda, J. J. (2012). *Procesamiento del Lenguaje Natural.* Univeridad Carlos III de Madrid.

Alemany, L. A. (2005). *Herramientas Libres para Procesamiento del Lenguaje Natural* . Jornadas Regionales de Sowtware Libre.

Gelbukh, A. (2010). Procesamiento de lenguaje natural. (S. d. Virtual, Ed.)

NLTK Project . (2015). *NLTK 3.0 Documentation* . Obtenido de http://www.nltk.org

Russell, S., & Norvig, P. (2004). *Inteligencia Artificial: Un enfoque moderno* (Segunda ed.). Pearson.

Pazos, A., & Barreiro, J. M. (1999). *Teleinformática.*

Sotomayor, J. (2006). *La Inteligencia Artificial en el Desarrollo de las Organizaciones Modernas.* Universidad del Valle de México .

Cañón, P. A., & Correa, S. R. (2011). *Procesamiento del lenguaje natural en la recuperación de información.* Univeridad de la Salle . Univeridad de la Salle .

VIcomtech. (2012). *Platafroma para el análisis de la ipinión de los conumidores y cuidadanos*. Obtenido de http://www.vicomtech.org/pr169/oferta-idi-proyecto-plataforma-para-el-analisis-de-la-opinion-de-los-consumidores-y-ciudadanos

Cortez, A., Vega, H., & Pariona, J. (2009). *Procesamiento de lenguaje natural.* Universidad Ricardo Palma . Universidad Ricardo Palma .

Russell, S. (1994). *Inteligencia Artificial: Un enfoque moderno.* Prentice Hall.

Contreras, H. Y. (2001). *Procesamiento del lenguaje natural basado en una gramática de estilos para el idioma español.* Univeridad de los Andes, Facultad de Ingeniería. Univeridad de los Andes.

Roldán, C. S. (2015). *Historia de Python*. Obtenido de https://www.codejobs.biz/es/blog/2013/03/03/historia-de-python

Python Software Foundation. (2015). *Python*. Obtenido de https://www.python.org/doc/

Marco, B. S. (2015). *Historia de Python*. Obtenido de http://www.mclibre.org/consultar/python/otros/historia.html

Soriano, G. (2009). *Introducción a Python.* Universidad de Buenos Aires.

Bird, S., & Loper, E. (2007). *NLTK: The Natural Language Toolkit* . University Of Melbourne . University Of Melbourne .

Jorge, M. (2011). *Historia de Twitter*. Obtenido de http://hipertextual.com/archivo/2011/03/historia-twitter/

Twitter Inc. (2015). *Twitter*. Obtenido de www.twitter.com

Martínez, J. (2012). *Guía para empezar en Twitter.*

Mollet, A., Moran, D., & Dunleavy, P. (2011). *El uso de Twitter en la investigación universitaria, la enseñanza y el impacto en las investigaciones: una guía para los académicos e investigadores.* . Universidad de León .

Twitter Inc. (2015). *Twitter Libraries*. Obtenido de https://dev.twitter.com/overview/api/twitter-libraries

Roesslein, J. (2009). *Tweepy Documentation*. Obtenido de http://tweepy.readthedocs.org/en/v3.2.0/index.html

Silberschatz, A., Korth, H., & Sudarshan, S. (2002). *Fundamentos de Bases de datos.* McGraw Hill.

Cruz, M. A. (2008). *Conceptos básicos de bases de datos.* Universidad Autónoma del Estado de Morelos. Universidad Autónoma del Estado de Morelos.

Martinez, G. (2008). *La historia del HTML (1989-2008).* Obtenido de http://aprendiendoweb.com/2008/08/la-historia-del-html-(1989-2008)

Menéndez, R. (2014). *Desarrollo Aplicaciones Web.* Universidad de Murcia.

The-PHP-Group. (2015). *PHP*. Obtenido de https://secure.php.net

Palomo, M., & Montero, I. (2010). *Programación en PHP a través de ejemplos.*

Vázquez, C. (2008). *Programación en PHP5. Nivel Básico.*

Sloman, A. (2007). *Artificial Intelligence. An Illustrative Overview*. The University of Birmingham. School of Computer Science.

Matrín, F., & Ruiz, J. L. (2012). *Procesamiento del lenguaje natural.*

Domiz, J., Roma, i. R., Fiadesio, C., Lantari, J., & Montesano, D. (2013). *Libro de Twitter: Conectados en 140 Caracteres.* Genes.